



## MÉTODOS DE CLUSTERIZAÇÃO EM CITOMETRIA DE FLUXO APLICADOS A IMUNOLOGIA

Elyr Teixeira de Almeida Alves

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Biomédica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Biomédica.

Orientadores: Marcio Nogueira de Souza

Alberto Felix Antonio da Nobrega

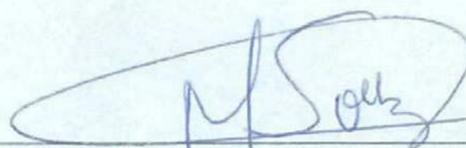
Rio de Janeiro  
Agosto de 2015

MÉTODOS DE CLUSTERIZAÇÃO EM CITOMETRIA DE FLUXO APLICADOS À  
IMUNOLOGIA

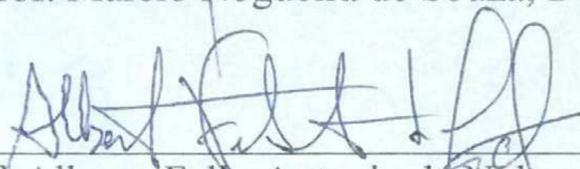
Elyr Teixeira de Almeida Alves

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ  
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA  
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS  
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM  
CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

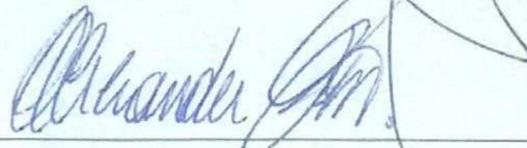
Examinada por:



Prof. Marcio Nogueira de Souza, D.Sc.



Prof. Alberto Felix Antonio da Nobrega, D.Sc.



Prof. Alexandre Visintainer Pino, D.Sc.



Prof. José Gabriel Rodríguez Carneiro Gomes, Ph.D.



Prof. Adriana Cesar Bonomo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

AGOSTO DE 2015

Alves, Elyr Teixeira de Almeida

Métodos de clusterização em citometria de fluxo aplicados à imunologia / Elyr Teixeira de Almeida Alves.

– Rio de Janeiro: UFRJ/COPPE, 2015.

XVI, 98 p.: il.; 29,7 cm.

Orientadores: Marcio Nogueira de Souza

Alberto Felix Antonio da Nobrega

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2015.

Referências Bibliográficas: p. 78-87.

1. Agrupamento de dados. 2.Clusterização. 3. Multidimensional. 4. Imunologia. I. Souza, Marcio Nogueira de *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

## DEDICATÓRIA

Dedico este trabalho a minha eterna companheira, namorada, noiva, esposa e amante, Luciana Brasil. Sem ela, não seria possível chegar até aqui “facilmente”. É incrível como o suporte de uma mulher pode fazer alguém ir longe. Não que não se possa ir longe sozinho. Mas quando você caminha bem acompanhado, a viagem é muito menos desgastante, além de uma probabilidade maior de sucesso. Cuidar da casa, trabalhar fora, cuidar das nossas duas princesas (Lavinia e Pérola), estar longe de seus pais, irmão e amigos e ainda me aguentar, definitivamente ela é um presente de Deus para mim. Pois, assim está escrito:

*“Toda mulher sábia edifica sua casa; mas a tola a derruba com as próprias mãos”*  
Provérbios 14:1

Aqui também faço uma menção ao meu amigo André Alvarenga. Quando, desde que iniciei no Inmetro ele sempre me incentivou e esteve ao meu lado. Suas palavras assertivas e sempre muito bem pensadas o faz um homem sábio e bondoso. Além dele ser incrivelmente bom no que faz. Agradeço a Deus por me tê-lo colocado em minha jornada.

Também dedico este trabalho ao meu sócio Daniel Morim. Em alguns momentos precisei me ausentar da empresa, ou de dividir tempo com ela, para me dedicar a este projeto. E se alguém merece esse reconhecimento por ter me ajudado em “cobrir os meus espaços”, esse é você. Que Deus te abençoe sempre.

Por fim, mas não menos importante, a toda a minha família. Pais, irmãos, familiares e amigos. Saí de Recife dizendo para meus pais que eu só iria fazer o mestrado no Rio de Janeiro, e que “rapidamente” (em dois anos) eu estaria de volta a minha cidade natal. Bom, os dois anos se transformaram em quase doze! Obrigado pela paciência de me esperar e compreender que eu ainda tinha que concluir outros objetivos. Com este trabalho, concluo mais uma audaciosa meta. Agora sim, posso voltar para casa.

A Deus seja dada toda honra, toda glória, pelos séculos e séculos. Amém.

## **AGRADECIMENTOS**

Meus agradecimentos se dirigem a todos que contribuíram com este trabalho.

Primeiramente ao meu orientador, professor Marcio Nogueira de Souza. Não apenas pela sua atitude em ensinar, disciplinar e corrigir, mas por entender que um aluno de pós-graduação é mais do que mão-de-obra especializada. Ele possui vida fora da universidade e compreender que a “vida lá fora” também faz parte da vida de um aluno faz com que um aluno admire e respeite seu orientador.

Professor Marcio, eu o admiro muito. Sua compreensão do todo o faz um professor completo. O Sr. tem o meu respeito e a minha total admiração.

Também fica aqui registrado meu agradecimento ao professor Alberto Nobrega. Nossas “longas noitadas” em seu laboratório foram muito produtivas e úteis para mim. Obrigado pelo seu tempo e extrema disponibilidade. Espero poder sempre retribuir toda a atenção despendida colaborando em novos trabalhos.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

## MÉTODOS DE CLUSTERIZAÇÃO EM CITOMETRIA DE FLUXO APLICADOS À IMUNOLOGIA

Elyr Teixeira de Almeida Alves

Agosto/2015

Orientadores: Marcio Nogueira de Souza

Alberto Felix Antonio da Nobrega

Programa: Engenharia Biomédica

Este trabalho apresenta duas contribuições para a melhoria de métodos de agrupamento de dados multidimensionais de citometria de fluxo; assim como a aplicação de um desses métodos em uma questão relevante para a área de imunologia, referente ao processo de maturação de células B. A primeira contribuição diz respeito ao chamado Método de agrupamento de dados com inicialização de centroides determinística baseada em modas. Este é formado pela mencionada técnica de inicialização de centroides em conjunto com o Critério de Informação Bayesiana (BIC), para estimação do melhor número de centroides, e pelo Modelo de Misturas Gaussianas (MMG), para modelagem e classificação. Este método foi avaliado com dados simulados e dados reais de citometria, e propiciou resultados melhores que aqueles obtidos com a inicialização aleatória das posições dos centroides em torno do vetor médio dos dados. A segunda contribuição diz respeito ao chamado Método Kappa Modificado, constituído da junção da técnica de inicialização de centroides aleatória no entorno do vetor médio dos dados, do Coeficiente Individual Kappa (CIK), para a escolha dos dados de maior relevância, e do MMG para modelagem e classificação de dados. O Método Kappa Modificado, não apenas classificou automaticamente um grupo de células de difícil seleção manual, como também permitiu que pela primeira vez este grupo celular pudesse ser associado às células da pré-zona marginal.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

## CLUSTERING METHODS IN FLOW CYTOMETRY APPLIED IN IMMUNOLOGY

Elyr Teixeira de Almeida Alves

August/2015

Advisors: Marcio Nogueira de Souza

Alberto Felix Antonio da Nobrega

Department: Biomedical Engineering

This work presents two contributions to the improvement of multidimensional data clustering methods in flow cytometry; as well as the application one of these methods in a relevant issue in the immunology area, referring to maturation process of cells B. The first contribution relates of data clustering method to initialize centroids based on deterministic mode. This is formed by mentioned centroids initialization technique in conjunction with the Bayesian Information Criterion (BIC) to better estimate the number of centroids, and the Gaussian Mixture Model (GMM) for modeling and classification. This method was evaluated with simulated and real data cytometry, and provided better results than those obtained with the random initialization of the positions of centroids around the middle of the data vector. The second contribution concerns the so-called Modified Kappa method, consisting of joint random centroids initialization technique surrounding the mean vector data, the Coeficiente Individual Kappa (CIK), to choose the most relevant data, and MMG to modeling and data classification. The Modified Kappa method, not just automatically classified a hard group of manual selection cells, but also allowed for the first time this group could be associated with cells of pre-zone marginal.

# SUMÁRIO

<b>DEDICATÓRIA .....</b>	<b>IV</b>
<b>AGRADECIMENTOS .....</b>	<b>V</b>
<b>LISTA DE FIGURAS .....</b>	<b>X</b>
<b>LISTA DE TABELAS .....</b>	<b>XIV</b>
<b>LISTA DE SIGLAS .....</b>	<b>XV</b>
<b>INTRODUÇÃO .....</b>	<b>1</b>
I.1 OBJETIVO.....	4
<b>FUNDAMENTOS TEÓRICOS .....</b>	<b>6</b>
II.1 CITÔMETRO DE FLUXO.....	6
II.1.1 <i>Visualização Gráfica</i> .....	9
II.2 FUNDAMENTOS DE IMUNOLOGIA.....	11
II.2.1 <i>Linfócitos B e T (células B e T) e os Desafios de suas Classificações</i> .....	11
II.2.2 <i>Células Imaturas, Foliculares e Zona Marginal</i> .....	12
II.2.3 <i>Desafios para Classificação</i> .....	13
II.2.4 <i>Designações Utilizadas na Expressão de Dados em Citometria de Fluxo</i> .....	13
II.3 ÍNDICE DE CONCORDÂNCIA ENTRE OBSERVADORES .....	14
II.3.1 <i>Coefficiente Kappa</i> .....	14
<b>ESTADO DA ARTE EM CITOMETRIA DE FLUXO.....</b>	<b>16</b>
III.1 AGRUPAMENTO DE DADOS EM CITOMETRIA DE FLUXO .....	16
III.2 FERRAMENTAS ACADÊMICAS .....	21
III.3 LIMITAÇÕES DA ANÁLISE CLÁSSICA.....	23
<b>ALGORITMOS ADAPTADOS .....</b>	<b>25</b>
IV.1 INICIALIZAÇÃO DE CENTROIDES BASEADA EM MODAS.....	25
IV.1.1 <i>Pseudocódigo da Inicialização dos Centroides</i> .....	31
IV.2 BAYESIAN INFORMATION CRITERION (BIC).....	33
IV.3 COEFICIENTE INDIVIDUAL KAPPA (CIK) .....	34
IV.4 MODELO DE MISTURA DE GAUSSIANAS (MMG).....	35
IV.4.1 <i>Distribuição Gaussiana</i> .....	36
IV.4.2 <i>A Mistura de Gaussianas</i> .....	37
IV.4.3 <i>Algoritmo Expectation-Maximization (EM) para Gaussianas</i> .....	38
IV.4.4 <i>Classificação dos dados a partir do MMG</i> .....	39
<b>MATERIAIS E MÉTODOS.....</b>	<b>40</b>

V.1 MÉTODO DE AGRUPAMENTO DE DADOS COM INICIALIZAÇÃO DE CENTROIDES DETERMINÍSTICOS BASEADO EM MODAS .....	40
V.2 MÉTODO DE AGRUPAMENTO DE DADOS QUASI-DETERMINÍSTICO BASEADO NO COEFICIENTE INDIVIDUAL KAPPA (CIK) OU MÉTODO KAPPA MODIFICADO.....	43
<b>RESULTADOS.....</b>	<b>45</b>
VI.1 MÉTODO DE AGRUPAMENTO DE DADOS DETERMINÍSTICOS BASEADO EM MODAS.....	45
<i>VI.1.1 Valores de BIC.....</i>	<i>46</i>
<i>VI.1.2 Influência dos Centroides Sugeridos e do Agrupamento Final de Dados .....</i>	<i>48</i>
<i>VI.1.3 Tempo Gasto e Iterações.....</i>	<i>51</i>
<i>VI.1.4 Concordância com o padrão ouro .....</i>	<i>52</i>
VI.2 MÉTODO DE KAPPA MODIFICADO PARA DEFINIÇÃO DA PRÉ-ZONA MARGINAL.....	53
<i>VI.2.1 Classificações Iniciais.....</i>	<i>53</i>
<i>VI.2.2 Pré-Zona Marginal.....</i>	<i>58</i>
<i>VI.2.3 Eixo Separador.....</i>	<i>60</i>
<b>DISCUSSÃO .....</b>	<b>63</b>
VII.1 DISCUSSÃO SOBRE O MÉTODO DE CENTROIDES BASEADOS EM MODAS.....	63
VII.2 DISCUSSÃO SOBRE O MÉTODO DE KAPPA MODIFICADO .....	65
<i>VII.2.1 Eixo Separador.....</i>	<i>70</i>
<i>VII.2.2 Clusters não Identificados .....</i>	<i>71</i>
<b>CONCLUSÃO.....</b>	<b>75</b>
<b>CONTRIBUIÇÕES CIENTÍFICAS RESULTANTES DESTE TRABALHO .....</b>	<b>77</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>	<b>78</b>
<b>APÊNDICE A – ALGUMAS DEFINIÇÕES MATEMÁTICAS .....</b>	<b>88</b>
<i>1 Cluster, Agrupamento de Dados, Classificação e Centroide .....</i>	<i>88</i>
<i>2 Medidas de Similaridade .....</i>	<i>89</i>
<i>3 Máxima Verossimilhança .....</i>	<i>90</i>
<b>APÊNDICE B – APRENDIZADO SUPERVISIONADO E NÃO-SUPERVISIONADO .....</b>	<b>92</b>
<b>APÊNDICE C – COEFICIENTE KAPPA .....</b>	<b>95</b>

## LISTA DE FIGURAS

- Figura 1 Diagrama esquemático de um citômetro de fluxo (WANDERLEY, 2006).
- Figura 2 Esquemático da câmara de focalização hidrodinâmica, com detalhe para a região de convergência (WANDERLEY, 2006).
- Figura 3 Gráficos típicos da área de citometria A) histograma de um parâmetro citométrico (Extraídos e adaptados de <http://www.biolreprod.org/content/69/4/1245/F6.expansion>), B) gráfico de espalhamento envolvendo dois parâmetros de um citômetro.
- Figura 4 Ilustração das intensidades luminosas das fluorescências para cada comprimento de onda de uma amostra de microesferas conhecidas (nuvem de pontos preta) e suas respectivas projeções ortogonais (Adaptado de ZENG, *et al.*, 2007).
- Figura 5 Modos de operação de quatro técnicas computacionais de redes neurais artificiais para uma única identificação de duas dimensões. a) *Perceptron* multicamadas; b) *Radial basis function*; c) *Asymmetric radial basis function*; d) Máquinas de vetores suporte (extraído e adaptado de BODDY, *et al.*, 2001).
- Figura 6 A) Principais vistas 2-D mais utilizadas na imunologia para detecção de células B da zona marginal, transicional e folicular; as imagens B) e C) representam duas abordagens para detecção das células B transicionais. D) apresenta o percentual de células transicionais (52%) que são encontradas na projeção IgM x CD21 pelo *gate* da Figura B) das células B transicionais desta Figura. E) apresenta o percentual de células transicionais (51%) que são encontradas na projeção CD24 x CD21 pelo *gate* da Figura C) das células B transicionais desta Figura.
- Figura 7 Gráficos de espalhamento de duas amostras de dados. A Figura do lado esquerdo é associada à amostra A e a Figura do lado direito à amostra B. Os eixos x e y são arbitrários.
- Figura 8 Do lado esquerdo a amostra A com a representação dos seus autovetores  $\mathbf{v1}$  e  $\mathbf{v2}$ . Do lado direito a mesma amostra com a projeção de seus pontos em seus autovetores. Os eixos x e y são arbitrários.
- Figura 9 Do lado esquerdo a amostra B com a representação dos seus autovetores  $\mathbf{v1}$  e  $\mathbf{v2}$ . Do lado direito a mesma amostra com a projeção de seus pontos em seus autovetores. Os eixos x e y são arbitrários.
- Figura 10 Histograma representativo de 3 modas. Os eixos x e y são arbitrários.
- Figura 11 Esquemático de varredura de picos de frequência para o lado direito a partir da moda principal. Os eixos x e y são arbitrários.

- Figura 12 Esquemático de varredura de picos de frequência para o lado esquerdo a partir da moda principal. Os eixos x e y são arbitrários.
- Figura 13 Contorno da distribuição Gaussianas em duas dimensões nas quais a matriz de covariância é a) geral, b) diagonal (alinhado ao eixo das coordenadas), c) proporcional a matriz identidade (extraído de BOTEV, 2006). Os eixos x e y são arbitrários.
- Figura 14 Representação gráfica dos conjuntos de dados. A) R15; B) Compound; C) Aggregation and D) Citometria FL7 x FL8.
- Figura 15 Apresentação gráfica dos valores de BIC (método clássico à esquerda e o método de moda à direita) em função do número de centroides para os quatro conjuntos de dados.
- Figura 16 Posições propostas dos centroides pela abordagem clássica do BIC (lado esquerdo) e do algoritmo proposto com base em modas (lado direito). Os centroides estão indicados por círculos vermelhos vazados. Para os gráficos referentes aos arquivos simulados (R15, Compound e Aggregation), os eixos x e y se referem a valores arbitrários para as coordenadas dos dados bidimensionais
- Figura 17 Resultado final do agrupamento para os conjuntos de dados simulados quando se usa a abordagem clássica do BIC (lado esquerdo) e do algoritmo baseado em modas (lado direito). Os eixos x e y se referem a valores arbitrários para as coordenadas dos dados bidimensionais.
- Figura 18 Resultado final do agrupamento para o conjunto de dados de citometria de fluxo realizada pelo especialista (acima), quando se usa a abordagem clássica (meio), e quando se usa o algoritmo baseado em modas (abaixo).
- Figura 19 Vistas canônicas 2-D das 300 mil células. Cada linha de imagens, após a primeira linha, representa os agrupamentos realizados pelo EFlow, indo de C1 a C6, referentes ao agrupamento do tipo A. Os *clusters* estão representados por pontos pretos.
- Figura 20 Vistas canônicas 2-D das 300 mil células. Cada linha de imagens, após a primeira linha, representa os agrupamentos realizados pelo EFlow, indo de C1' a C6', referentes ao agrupamento do tipo A'. Os *cluster* estão representados por pontos pretos.
- Figura 21 A Figura do lado esquerdo apresenta as células do C2' em sua integridade, com 32.476 células. Na Figura do lado direito a mesma vista, CD24 x CD21, apresenta o mesmo *cluster* mas com uma redução 11,06%.
- Figura 22 A Figura A apresenta um histograma do total de células (em valores percentuais) para cada um dos 6 *clusters* dos grupos C e C'. A Figura B informa a localização dos *clusters* C2'a e C2'b nas vistas CD24 x CD21 com pontos na cor preta. A Figura C apresenta um histograma do total de células (em valores percentuais) dos grupos C e C' com destaque para o novo subgrupo 2b do grupo C'.

- Figura 23 A Figura A apresenta o *cluster* C2' original, e na Figura B os dois *clusters* formados após a etapa de agrupamento do algoritmo de MMG nas vistas CD24 x CD21.
- Figura 24 As Figuras do lado esquerdo representam as projeções dos *clusters* C1', C3'+C4' e C2'a, nas vistas a) CD24 x CD21, b) CD23 x CD21 e c) Eixo 1 x Eixo 2 determinadas empiricamente. Os *clusters* C1', C3'+C4' e C2'a representam as células transiticionais (azul), folicular (preto) e zona marginal (vermelho), respectivamente. As Figuras do lado direito denotam os mesmos *clusters* e as projeções do lado esquerdo, mas agora representados em curvas de nível.
- Figura 25 Apresentação dos *clusters* C1', C3'+C4' e C2'a com o(s) *cluster(s)*: (A) C2'b; (B) C5'; (C) C2'b e C5'. As Figuras A, B e C são representadas por curvas de nível. A Figura D por densidade de pontos.
- Figura 26 Nas vistas IgM x IgD são apresentadas as 300 mil células com destaque na cor vermelha para os grupos: Transicional; Zona Marginal; e *cluster* C2'b, nesta ordem.
- Figura 27 Nas vistas CD24 x CD21 são apresentadas as 300 mil células com destaque para: (A) *cluster* C2'b em vermelho; e (B) células Transicionais em verde.
- Figura 28 Os pontos vermelhos representam as células classificadas no *cluster* C2'b na vista CD23 x CD21 em relação ao total de células.
- Figura 29 *Gate* padrão para classificação das populações de Zona Marginal e Pré-Zona Marginal, Foliculares e células Transicionais.
- Figura 30 A Figura A apresenta a vista IgM x IgD com destaque para o *gate* em células com a configuração IgM<sup>high</sup> e IgD<sup>low</sup>. A Figura B apresenta as células selecionadas na Figura anterior (A) nas vistas CD24 x CD93. Na Figura C (em duas Figuras menores) podem ser vistas as células que foram selecionadas no *gate* anterior (Figura B) nas vistas CD23 x CD21.
- Figura 31 Com a projeção nas vistas CD23 x CD21 as três imagens apresentam os possíveis comportamentos das células a partir das células B transiticionais até a zona marginal. A Figura A demarca o caminho clássico dos grupos: transiticionais (i), pré-zona marginal (ii) e a zona marginal (iii). A Figura B apresenta um possível caminho de maturação da transiticional (i) para o *cluster* C2'b (iv) e em seguida para a zona marginal (iii). A Figura C apresenta uma segunda possibilidade partindo da região das células transiticionais (i), chegando na região do *cluster* C2'b (iv) e dela se direcionando para a pré-zona marginal (ii) para então chegar na região da zona marginal (iii).

- Figura 32 A Figura A apresenta a vista IgM x CD21 com um *gate* na região determinada pela zona marginal e pré-zona marginal. A Figura B apresenta as células *gateadas* na Figura anterior com a separação da zona marginal (à esquerda) e da pré-zona marginal (à direita) nas vistas CD23 x CD21. A Figura C apresenta a distribuição dessas células nos *clusters* C5', C4' e C2'b nas vistas IgM x CD21, com 59%, 20% e 12% respectivamente
- Figura 33 Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista CD24 x CD21.
- Figura 34 Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista CD23 x CD21.
- Figura 35 Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista IgM x IgD.

## LISTA DE TABELAS

Tabela 1	Características dos conjuntos de dados utilizados na avaliação.
Tabela 2	Número de centroides reais em comparação com as quantidades de centroides propostos pelo método do BIC clássico e pelo método baseado por modas com o BIC.
Tabela 3	Número de iterações e tempo gasto para a etapa de agrupamento para o método de BIC clássico e para o método baseado em modas.
Tabela 4	Comparação dos coeficientes de Kappa (K) entre o método clássico do BIC e o método baseado em modas. Ambos utilizando a técnica de MMG para seu agrupamento.
Tabela 5	Apresentação do total de células em cada <i>cluster</i> para os agrupamentos A e A' de uma classificação de cada tipo escolhida ao acaso, bem como sua distribuição percentual em relação ao total de 300 mil células.
Tabela 6	Apresenta o quantitativo original de células, o quantitativo de células após o uso do CIK e o percentual de células reduzidas, para ambos os agrupamentos A e A'.

## LISTA DE SIGLAS

7-AAD	7 aminoactinomicina D
Anti-220B	Anticorpo 220B
Anti-CD21	Anticorpo CD21
Anti-CD23	Anticorpo CD23
Anti-CD24	Anticorpo CD24
Anti-CD93	Anticorpo CD93
Anti-IgM	Anticorpo IgM
APP-CY7	Fluorocromo Aloficocianina Cy Chrome 7
BIC	Bayesian Information Criterion
C57BL	Tipo de camundongo
CD	Cluster de Diferenciação
CFH	Câmara de Focalização Hidrodinâmica
CIK	Coefficiente Individual Kappa
COPPE	Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
DECH	Doença de Enxerto Contra-Hospedeiro
ECD	Ficoeritrina-Texas Red
EM	Expectation Maximization
FACS	Fluorescence-Activated Cell Sorting
FITC	Fluorocromo Isoticionato de Fluoresceína
FO	Folicular
FSC	Forward Scatter
GMM	Gaussian Mixture Model
HT	High Throughput
IgD	Imunoglobulina D
IgM	Imunoglobulina
MMG	Modelo de Mistura de Gaussianas
PE	Fluorocromo Ficoeritrina
PERCP	Fluorocromo Proteína Clorofilperidina
PI	Iodeto de propídio

PZM	Pré-Zona Marginal
RBS	Radial Basis Function
ROI	Region of Interest
SSC	Side Scatter
SVM	Support Vector Machine
TR	Transicional
UFRJ	Universidade Federal do Rio de Janeiro
ZM	Zona Marginal

# CAPÍTULO 1

## INTRODUÇÃO

A citometria de fluxo é uma técnica especializada em contar, examinar e classificar grande volume de partículas microscópicas, quando suspensas em um fluido. Este tipo de técnica permite a análise simultânea de parâmetros físicos, químicos e biológicos; sendo uma técnica que permite precisão, velocidade e repetibilidade (JONKER, *et al.*, 2000; WILKINS, *et al.*, 2001; KOTSIANTIS, 2007).

Na área médica esta técnica tem encontrado uma grande gama de aplicações e se caracteriza como uma ferramenta para a análise qualitativa e quantitativa de dados biológicos. Dentre tais aplicações podem ser citadas: análise em hematologia veterinária (NAKAGE, *et al.*, 2005) para a contagem celular, fórmula leucocitária, análise das populações celulares da medula óssea (MO) (MODESTO, *et al.*, 2006); oncologia (CAVALCANTI JÚNIOR, *et al.*, 2004) no diagnóstico e prognóstico, monitoramento, tratamento do câncer; imunologia (BERTHO, *et al.*, 2000) em estudos de subpopulações de células T e B, fenotipagem, ativação linfocitária, apoptose; em farmacologia no estudo de cinética celular e resistência a drogas e sobre a eficácia de novos medicamentos quando administrados em baixas concentrações (SMITH, *et al.*, 2007); na microbiologia em diagnóstico bacteriano e viral, sensibilidade a antibióticos, na genética para estudo de cariótipos, diagnóstico de portador de AIDS, diagnóstico pré-natal e fenotipagem para transplantes, dentre outras.

Apesar da citometria estar difundida no meio médico, a análise dos dados brutos gerados pelos citômetros, equipamentos usados para a realização dos experimentos de citometria, continua sendo uma preocupação da área de pesquisa. Por exemplo, um dos objetivos dos sistemas de análise em desenvolvimento é o de classificar corretamente grupos de populações celulares similares. Tal classificação se torna bastante complexa usando-se somente as tradicionais ferramentas baseadas em identificação visual, mesmo

para um operador treinado. Quando o número de dimensões brutas do citômetro é reduzida, a análise visual pode ser prática e rápida; entretanto, pode-se perder informação significativa por causa dessa diminuição, tais como: tamanho da célula, granulosidade etc.

Com um número cada vez maior de parâmetros fornecidos pelos citômetros modernos, a classificação de células pode se tornar mais acurada, mas, por outro lado, a análise realizada em tal distinção celular torna-se também cada vez mais complexa, quando não inviável para a inspeção visual de um operador humano. Ainda neste sentido, dependendo da aplicação, somente com um número maior de parâmetros é que se consegue obter informação necessária à identificação de certas populações celulares. Assim, metodologias automáticas para o agrupamento de populações de células semelhantes têm sido pesquisadas (WILKINS, *et al.*, 2001). Tais técnicas automáticas acarretam redução do tempo de análise e maior confiabilidade e reprodutibilidade da análise celular. Uma vantagem é combinar algumas dessas técnicas com computação paralela utilizando placas gráficas aceleradoras, reduzindo o tempo de análise e de processamento de dados (PANGBORN, 2009).

Apesar dos vários estudos existentes, nenhuma técnica computacional de separação ou classificação de dados pode ser considerada como definitiva na área de citometria de fluxo, isto é, não há algoritmo computacional que solucione todas as possíveis situações experimentais reais. Uma vez que cada técnica computacional possui suas próprias particularidades, o emprego de uma técnica isolada não garante, muitas vezes, um resultado satisfatório (PANGBORN, 2009; BISHOP, 2006). Em alguns casos, é necessário o emprego combinado de duas ou mais técnicas de classificação para maximização de um resultado. Entende-se aqui o termo maximização de um resultado, como a capacidade de realizar um agrupamento o mais assemelhado possível com o agrupamento real dos dados.

Não obstante o desenvolvimento de um número cada vez maior de técnicas computacionais em citometria, o foco dessas se concentra principalmente na padronização da análise de classificação dos tipos celulares em clusters característicos, o que já é realizado com sucesso pelo operador humano. Individualmente, essas análises não apresentam grande dificuldade, pois se limitam a identificar clusters muito bem delineados. Todavia, existe uma grande variabilidade na descrição desses clusters, que é por muitas vezes agravada pela insuficiência da formação técnica do operador. Essas

dificuldades precisam ser superadas para que se possa estabelecer uma base de dados universal, útil ao diagnóstico e práticas médicas. É, portanto, no sentido de obter uma padronização automática daquilo que o operador humano já realiza, que a maioria dos desenvolvimentos de técnicas computacionais em citometria têm sido feitos. Entre estes podem ser citados: as Redes Neurais Artificiais (BODDY, *et al.*, 1994; GODAVARTI, *et al.*, 1996; WILKINS, *et al.*, 1999; AL-HADDAD, *et al.*, 2000; JONKER, *et al.*, 2000; BODDY, *et al.*, 2001; MORRIS, *et al.*, 2001; WILKINS, *et al.*, 2001; WUNSCH e XU, 2005; QUINN, *et al.*, 2007; WELLER, *et al.*, 2007), Lógica *Fuzzy* (WILKINS, *et al.*, 2001; WUNSCH e XU, 2005), Máquinas de Vetores de Suporte (*Support Vector Machines – SVM*) (MORRIS, *et al.*, 2001; BODDY, *et al.*, 2001; SVENSSON, *et al.*, 2014), *Hierarchical Clustering* (WUNSCH e XU, 2005; KITSOS, *et al.*, 2007) e *K-means* (WUNSCH e XU, 2005; ZENG, *et al.*, 2007).

Ao lado da questão da padronização da análise de subdivisão em clusters característicos, as técnicas computacionais em citometria podem contribuir de forma mais sutil e complexa, como instrumentos para exploração da estrutura de dados multiparamétricos complexos. Aqui, os algoritmos contribuem para desvendar a própria estrutura dos dados, e não para padronizar uma classificação já estabelecida por um operador humano. Essa outra aplicação de algoritmos computacionais em citometria se destina principalmente ao trabalho de pesquisa, no qual se busca identificar correlações entre populações celulares diversificadas e que mantêm entre si uma intrincada rede de relações biológicas. Como exemplo típico, podemos citar o processo de diferenciação das células hematopoiéticas a partir de progenitores que se originam da proliferação e diferenciação progressiva das células tronco. A caracterização exata desse processo passa pela identificação de cerca de uma dezena de tipos celulares intermediários, normalmente realizada com o auxílio de citometria multiparamétrica, com 8 ou mais parâmetros.

Um problema relevante no contexto da pesquisa biológica atual consiste na caracterização dos tipos celulares intermediários que pertencem ao processo de maturação dos linfócitos B (ALLMAN *et al.*, 2004). Os linfócitos B, ou células B, são formados na medula óssea a partir de progenitores derivados das células tronco hematopoiéticas. As células B recém-formadas se apresentam como uma população funcionalmente imatura, incapaz de responder a uma série de estímulos biológicos e de adquirir a competência para produção de anticorpos. A maturação funcional dessas células, que se tornam competentes para produção de anticorpos, vai ocorrer no baço, que é um importante órgão

linfoide. Neste órgão, as células B imaturas completam seu processo de diferenciação, gerando duas populações de linfócitos B maduros: células B foliculares (FO) e células B de zona marginal (ZM). Esses dois subtipos diferem em várias funções biológicas, tais como resposta acelerada ao estímulo por determinadas moléculas de origem viral ou bacteriana e a capacidade de proliferação celular e produção de anticorpos.

No processo de maturação dos linfócitos B no baço, um ponto central consiste na decisão que cada célula faz, ao tornar-se uma célula folicular ou de zona marginal. Essa opção não é aleatória e depende crucialmente da sinalização molecular resultante da atuação das moléculas imunoglobulinas que estão na membrana do linfócito. Como cada linfócito imaturo expressa um rearranjo distinto de imunoglobulina, cada célula, por assim dizer, toma sua própria decisão. Nesse processo, as células B imaturas passam por estados intermediários de maturação, onde de fato ocorrem os processos biológicos celulares de maturação.

Tendo em vista o cenário acima mencionado, torna-se importante caracterizar os estados intermediários e quantificá-los para avaliar se os processos de maturação estão ocorrendo normalmente. A identificação das populações intermediárias não é simples, pois estas células apresentam um fenótipo híbrido, entre células imaturas e maduras (ALLMAN e PILLAI, 2008). Essas populações não formam clusters bem definidos e sua classificação necessita de citometria multiparamétrica, com a utilização de 6 ou mais marcadores. Essa análise apresenta maior complexidade, pois depende do acompanhamento simultâneo da variação da expressão de várias moléculas na superfície celular. Esse tipo de análise coloca um novo desafio para as técnicas computacionais, que é a descrição fidedigna da estrutura dos dados num espaço multidimensional.

## **I.1 Objetivo**

O presente trabalho investiga e desenvolve métodos de agrupamento de dados multidimensionais aplicáveis na área de citometria de fluxo, objetivando contribuir com a identificação de populações de células do sistema imunológico, especificamente para a área de imunologia tumoral.

Para alcançar o objetivo principal, foram delineados os seguintes objetivos parciais:

- Implementar, testar, avaliar e validar técnicas de agrupamento de dados de pequena dimensionalidade (até 3 dimensões)
- Estender as técnicas de agrupamento de dados de baixa dimensionalidade para dados citométricos que envolvam mais de 3 parâmetros de fluorescência;
- Aprimorar e testar métodos de agrupamento num problema com relevância biológica.

Este trabalho está dividido da seguinte forma: no Capítulo 2 são abordados os fundamentos teóricos necessários para o entendimento dos métodos e das ferramentas empregadas. No Capítulo 3 é descrito o estado da arte, apontando as técnicas de agrupamentos de dados ao longo do tempo com o auxílio da academia. Uma seção deste capítulo apresenta as principais limitações encontradas por citometrista quando necessitam classificar dados multidimensionais de citometria de fluxo. O Capítulo 4 apresenta as duas contribuições realizadas neste trabalho para problemas de agrupamentos de dados. No Capítulo 5 são apresentados os materiais e métodos utilizados, assim como os seus critérios para a etapa de avaliação dos resultados. O Capítulo 6 são apresentados os resultados das duas contribuições desenvolvidas, enquanto no Capítulo 7, são discutidas questões relacionadas ao emprego das mesmas. O último capítulo, Capítulo 8, apresenta a conclusão deste trabalho, aponta vantagens e limitações, assim como sugestões para trabalhos futuros.

# **CAPÍTULO 2**

## **FUNDAMENTOS TEÓRICOS**

Como este trabalho é multidisciplinar, este capítulo se destina a fornecer alguns fundamentos necessários para uma melhor compreensão das questões abordadas no texto. Nele são apresentados alguns conceitos da área de ciências biológicas e outros da área de ciências exatas.

### **II.1 Citômetro de Fluxo**

A caracterização de células inclui, entre outras, informações relativas ao seu tamanho e complexidade interna. A determinação de tais características pode ser feita com um arranjo experimental composto por uma câmara de focalização, lasers, lentes e fotodetectores. Tal conjunto é a base do equipamento denominado citômetro de fluxo e que se encontra ilustrado, numa das suas possíveis realizações, na Figura 1 (WANDERLEY, 2006).

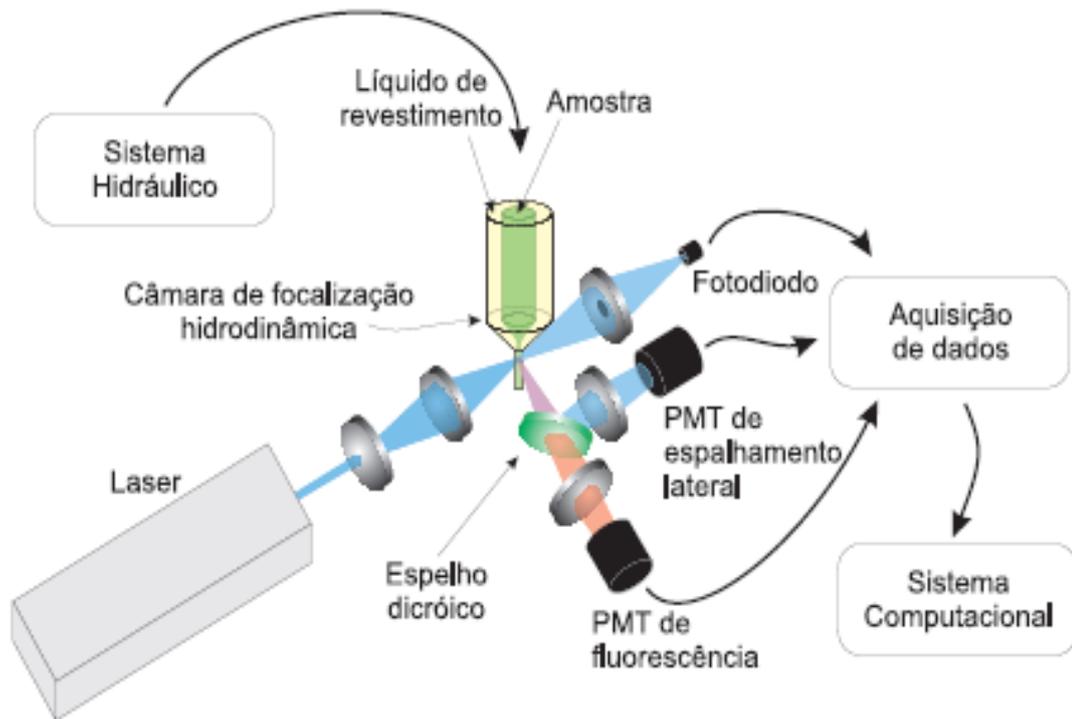


Figura 1. Diagrama esquemático de um citômetro de fluxo (WANDERLEY, 2006).

Uma importante parte de um citômetro de fluxo é a sua câmara de focalização hidrodinâmica (CFH). Esta câmara tem a finalidade de enfileirar individualmente as células da amostra, com o intuito de se avaliar individualmente cada célula (ORMEROD, 2000; SHAPIRO, 2005).

Uma ideia simples para enfileirar essas células seria a confecção de um capilar com um diâmetro semelhante ao diâmetro das células analisadas. Assim, se poderia garantir que apenas uma célula conseguiria atravessar o capilar por vez. Entretanto, para células maiores, o entupimento do capilar inviabilizaria um novo estudo, o que torna esta metodologia não aplicável na prática. Para eliminação deste problema (entupimento) Crosland-Taylor (1953) desenvolveu um sistema que é chamado atualmente de câmara de focalização hidrodinâmica. Este dispositivo possui duas partes concêntricas, em que a menor é responsável por conduzir a amostra até uma região de convergência com a parte maior, a qual conduz um líquido chamado de líquido de revestimento. A Figura 2 ilustra uma possível realização de uma CFH.

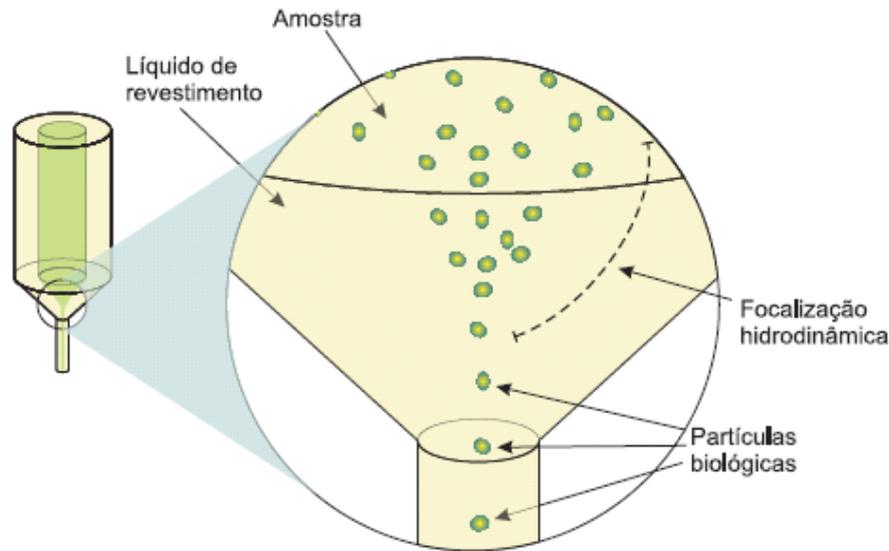


Figura 2. Esquemático da câmara de focalização hidrodinâmica, com detalhe para a região de convergência (WANDERLEY, 2006).

O princípio de funcionamento da CFH baseia-se no princípio de *Bernoulli*, onde devido à redução gradual do diâmetro na área de convergência, as células se movem para regiões de baixa pressão quanto maior for a velocidade do fluxo, desde que possa ser garantido um fluxo laminar (RATCOM, *et al.*, 1984; WANDERLEY, 2006).

A partir do momento que a amostra encontra o líquido de revestimento, as células serão conduzidas, uma a uma, à excitação luminosa de um laser. Após a iluminação individual das células pelo laser, a luz incidente é redirecionada em várias direções, sendo que a intensidade luminosa espalhada na mesma direção da fonte de luz de excitação é chamada de Espalhamento Direto ou Frontal (*Forward Scatter* ou FSC). Por outro lado, a quantidade de luz espalhada na direção perpendicular à direção de excitação é chamada de Espalhamento Lateral (*Side Scatter* ou SSC). O FSC se relaciona com o tamanho da partícula, enquanto o SSC com a granulosidade ou complexidade celular (CROSLAND-TAYLOR, 1953).

Outra forma de se obter mais informações sobre a célula (estrutura, função e vitalidade, por exemplo) é realizando sua marcação com fluorocromos, antes de sua passagem pelo citômetro. Fluorocromos são substâncias que absorvem energia luminosa num dado comprimento de onda e que emitem essa energia luminosa num comprimento de onda maior, sendo este, visível. Quando um fluorocromo é excitado, a resposta a essa excitação é a emissão de uma fluorescência, e considerando que ele está ligado a um anticorpo ou a uma molécula específica, a detecção da quantidade de luz de comprimento

de onda da fluorescência está relacionada à existência das moléculas que o fluorocromo foi projetado para detectar. Quanto maior o número de canais de fluorescência num citômetro, maior a possibilidade de realização de análises mais complexas de tipo e funções das várias células que compõem a amostra. Os fluorocromos mais conhecidos são: o FITC, PE, ECD, PerCP, PI, 7-AAD, iodeto de propídeo, laranja de acridina, pironina Y, fluo-3, entre outros (BERTHO, 2001).

Todos esses parâmetros (FSC, SSC e canais de fluorescência) são chamados de parâmetros brutos ou primários de um citômetro de fluxo e devem ser mensurados, armazenados e analisados por um sistema de controle e análise. Uma vez que um citômetro de fluxo é composto por sistemas hidráulicos, pneumáticos, opto-eletrônicos e computacionais, é necessário se ter uma base multidisciplinar para o entendimento e desenvolvimento deste tipo de equipamento.

### II.1.1 Visualização Gráfica

Existem várias formas de se representar os dados primários fornecidos por um citômetro de fluxo. Quando apenas um parâmetro é avaliado, a representação gráfica mais usual é um histograma que expressa o número de células ou partículas para cada valor do parâmetro em questão (Figura 3A). Quando dois parâmetros são analisados simultaneamente (FSC e SSC, por exemplo) uma representação alternativa é o gráfico de espalhamento (FRELINGER, *et al.*, 2010), como na Figura 3B. Ambos os gráficos possuem escala arbitrária de intensidade luminosa para os eixos que se referem aos canais.

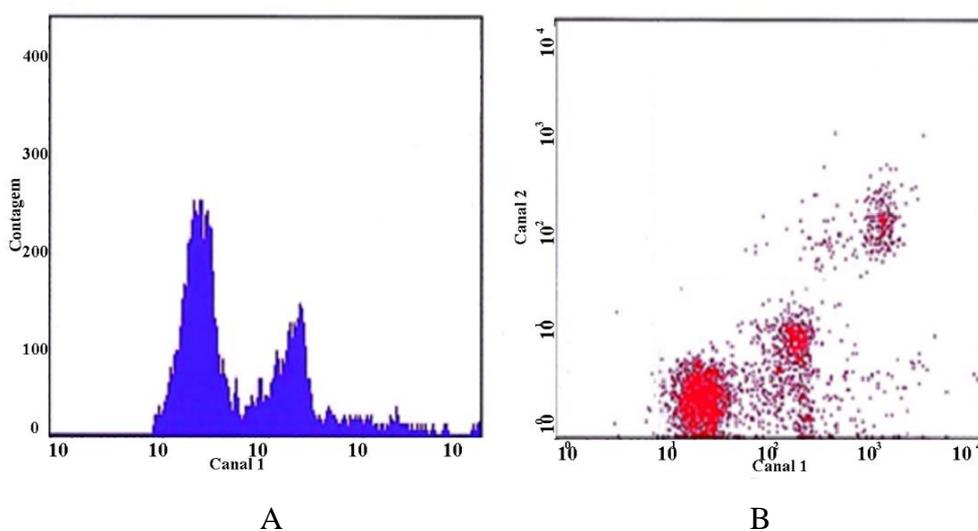


Figura 3. Gráficos típicos da área de citometria A) histograma de um parâmetro citométrico (Extraídos e adaptados de <http://www.bioreprod.org/content/69/4/1245/F6.expansion>), B) gráfico de espalhamento envolvendo dois parâmetros de um citômetro.

A representação de partículas com três parâmetros poderia ser feita por um gráfico tridimensional, que indicaria cada tipo de população como uma nuvem de pontos (Figura 4). Entretanto, a forma de se analisar esses dados a partir de um gráfico 3-D não é simples, pois o mesmo sempre é visualizado em um monitor bidimensional, dificultando a interpretação de um observador. Por isso, muitas vezes, a solução encontrada é recorrer a projeções ortogonais bidimensionais para que com uma análise individual de cada plano se possa ter uma melhor ideia das distribuições das diferentes partículas. A Figura 4 ilustra uma amostra de microesferas de dimensões conhecidas e suas projeções nos planos formados pelos três dados de fluorescência (ZENG, *et al.*, 2007). As unidades da Figura são arbitrárias e representam a intensidade luminosa das fluorescências para cada comprimento de onda.

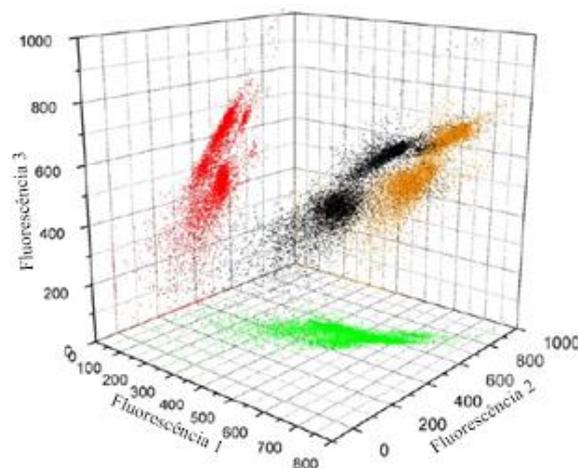


Figura 4. Ilustração das intensidades luminosas das fluorescências para cada comprimento de onda de uma amostra de microesferas conhecidas (nuvem de pontos preta) e suas respectivas projeções ortogonais (Adaptado de ZENG, *et al.*, 2007).

Devido ao fato de, em citometria de fluxo, se analisar dados multiparamétricos, a informação necessária à separação adequada de populações de células pode estar escondida na alta dimensionalidade. Isto leva, normalmente, o especialista a realizar seleções manuais de determinados parâmetros citométricos para visualizações bidimensionais, também chamadas de *gates* (CHEN, *et al.*, 2015), até que ele possa chegar ao conjunto de dados desejado. Esse trabalho é exaustivo, suscetível a erros e pouco reproduzível, por se tratar de um procedimento puramente manual e visual, o qual depende exclusivamente da experiência do usuário.

## II.2 Fundamentos de Imunologia

### II.2.1 Linfócitos B e T (células B e T) e os Desafios de suas Classificações

O linfócito é um tipo de leucócito (células sanguíneas brancas ou glóbulos brancos) e são fabricados pelas células-tronco linfoides presentes na medula óssea vermelha. Os linfócitos deixam a medula óssea, circulam o sistema linfático e sanguíneo e se localizam em vários órgãos linfoides. Devido à produção e apresentação de receptores de superfície celular que se ligam aos antígenos, os linfócitos medeiam os atributos imunológicos definidos como: especificidade, diversidade, memória e reconhecimento do próprio e do não próprio (KINDT, *et al.*, 2008).

Eles constituem uma família de células esféricas, com um diâmetro que varia de 6 a 8 µm, sendo que os linfócitos que possuem esta dimensão são chamados de linfócitos pequenos, os linfócitos B e T. Já no sangue circulante ocorre uma pequena porcentagem de linfócitos maiores, que podem chegar até 18 µm de diâmetro, também conhecido pelo nome de células exterminadoras naturais (*Natural killer*). Em indivíduos saudáveis, os linfócitos representam de 20% a 30% dos leucócitos sanguíneos, e possuem um papel importante na defesa do organismo, possuindo receptores capazes de identificar invasores, sendo que cada um deles possui suas particularidades (JUNQUEIRA e CARNEIRO, 2004).

As duas principais populações de linfócitos são: linfócitos B (células B) e linfócitos T (células T). Embora ambos os linfócitos B e T se originem na medula óssea, a maturação dos linfócitos B ocorre na medula óssea, enquanto que a maturação dos linfócitos T ocorre no timo (glândula situada no tórax, atrás do osso esterno e que faz parte do sistema imunológico).

As células B representam 5% a 10% dos linfócitos do sangue. Elas penetram nos vasos sanguíneos através de movimentação ameboide, sendo transportadas pela circulação até alcançarem os órgãos linfáticos, onde se alojam, com exceção do timo. Os linfócitos B produzem os anticorpos sanguíneos. Quando estas células são ativadas por antígenos, proliferam e se diferenciam em plasmócitos (células produtoras de anticorpos) (JUNQUEIRA e CARNEIRO, 2004).

Já no timo, os linfócitos T se diferenciam em células *T helper*, T supressora e T citotóxica. Os *T helper* estimulam a transformação dos linfócitos B em plasmócitos. Os

linfócitos T supressores inibem a resposta humoral e celular e apressam o término da resposta imunitária. Os linfócitos T citotóxicos agem diretamente sobre as células estranhas e infectadas por vírus através de dois mecanismos: a produção de proteínas, que abrem orifícios nas membranas plasmáticas, provocando a lise<sup>1</sup> das células, e através da indução das células-alvo a entrarem em apoptose<sup>2</sup>. Estas células representam de 65% a 75% dos linfócitos sanguíneos (JUNQUEIRA e CARNEIRO, 2004).

## II.2.2 Células Imaturas, Foliculares e Zona Marginal

As células B imaturas são formadas na medula óssea e amadurecem em locais periféricos secundários, predominantemente no baço, gerando duas subpopulações histológica, fenotípica e funcionalmente distintas: células foliculares e as células da zona marginal (KUMARARATNE, *et al.*, 1981; MARTIN e KEARNEY, 2002).

As células B foliculares são pequenos linfócitos de recirculação (quando este retorna à corrente sanguínea várias vezes), com uma baixa expressão de IgM e com uma alta expressão de IgD (o prefixo Ig significa imunoglobulinas), formando os folículos das células B na polpa branca (localizada no baço). Em comparação, as células B da zona marginal são identificadas como células de tamanho médio, com alta expressão de IgM e com baixa expressão de IgD, residentes na zona marginal do baço (KUMARARATNE, *et al.*, 1981). As células B da ZM são maioria na ZM, juntamente com as células mieloide e dendríticas. Embora a ZM seja um local de passagem para reentrada de células B e de várias células B já ativadas (LIU *et al.*, 1988), a maior parte das células B encontradas na ZM são células que nunca encontraram um antígeno diferente (KUMARARATNE e MACLENNAN, 1982). A estas células é dado o nome de *naïve*. Tais células B *naïve* apresentam características fenotípicas distintas, tais como alta expressão de CD21 (KUMARARATNE *et al.*, 1981; MARTIN e KEARNEY, 2002) e rápida habilidade para responder a certos antígenos, ou patógenos, filtrados da corrente sanguínea (GUINAMARD, *et al.*, 2000; MARTIN e KEARNEY, 2002). Essas características sugerem que as células B da ZM desempenham funções especializadas, as quais atuam em sinergia com a imunidade adaptativa, que é inicializada pelas células B FO em conjunto com as células T (MARTIN e KEARNEY, 2002).

---

<sup>1</sup> Lise é o processo de ruptura da membrana plasmática que leva à morte da célula e à liberação de seu conteúdo.

<sup>2</sup> Apoptose é a morte programada da célula.

### II.2.3 Desafios para Classificação

O uso de dados multiparamétricos por citometria de fluxo desempenha um papel central na caracterização fenotípica de células hematopoiéticas. As maiores linhagens de células hematopoiéticas são identificadas pela expressão de uma combinação definida de moléculas de superfície. Por exemplo, as células T são caracterizadas pela expressão de receptores CD3/ $\alpha\beta$  ou CD3/ $\gamma\delta$ ; CD19 é um marcador canônico para linhagem de células B; CD11b identifica células mieloides maduras, as quais podem ser subdivididas em diferentes tipos de células baseadas na expressão adicional de células de superfície, tais como: Ly6G, Ly6C, CD115, CD11c, entre outras. Enquanto a maior parte da linhagem de células hematopoiéticas são fielmente identificadas com marcadores bem estabelecidos no meio científico, este comportamento não é facilmente encontrado na fenotipagem de diferentes subpopulações de uma dada linhagem, por exemplo, nas subpopulações de macrófagos, células dendríticas ou células B.

A dificuldade para caracterizar subpopulações de células com uma dada linhagem está frequentemente relacionada com o espectro contínuo da expressão das moléculas de superfície que são utilizadas como marcadores para diferenciar subtipos de células. Como visto anteriormente, os fluorocromos se ligam a moléculas específicas de uma superfície celular, e como a sua intensidade luminosa produzida pela fluorescência é contínua, torna-se difícil analisar e classificar a distribuição de várias células diferentes ao mesmo tempo (HERZENBERG, *et al.*, 2006).

### II.2.4 Designações Utilizadas na Expressão de Dados em Citometria de Fluxo

Em citometria de fluxo é comum o emprego de rótulos (*labels*) para diferenciar a expressão dos dados contidos em um gráfico bidimensional, mesmo que esses *labels* não tenham um consenso universal sobre sua aplicação (padronização). Os *labels* mais frequentemente utilizados são: *low*, *high*, *bright*, positivo, negativo e até mesmo sinais matemáticos (+), (++) , (-) e (- -). Uma breve explicação esclarece seu significado para os leitores mais leigos.

Populações celulares são usualmente definidas por símbolos (+) ou (-) para indicar se elas contêm (ou não) uma determinada molécula. Por exemplo: a célula “CD34+, CD31-” expressa a presença da molécula CD34 e ausência da molécula CD31. O *label low* representa que há uma baixa quantidade de um determinado marcador, mas que mesmo assim ele é presente. Já o *label high* representa uma alta quantidade de um

marcador. O *label bright* expressa uma quantidade ainda maior de determinado marcador. Sinônimos para esta situação são também usados, tais como: *low*, *dim*, (-) e (- -) para representar o *low* e *high*, *bright*, (+) e (+ +) para *high* (MESKAS e DROUMÉVA 2015).

## II.3 Índice de Concordância Entre Observadores

Quando não há uma classificação exatamente aceita como padrão para determinada observação experimental é comum se submeter tal observação a vários observadores e gerar uma métrica que avalie o grau de concordância entre as classificações realizadas pelos mesmos. Neste trabalho a concordância entre especialistas, sejam eles humanos ou computacionais, foi avaliada pelo Coeficiente Kappa, o qual é descrito a seguir.

### II.3.1 Coeficiente Kappa

O coeficiente Kappa é uma medida de concordância estatística entre classificações de um mesmo evento realizadas por diferentes observadores, tendo sido formulada por JACOB COHEN em 1960. Ela pode ser aplicada para observadores humanos ou não. Esta medida de concordância tem como valor máximo 1, o qual significa concordância total entre as classificações dos observadores. O valor zero significa classificações completamente não correlacionadas, ou ao acaso; enquanto valores negativos, até o limite de -1, significam discordância entre as classificações dos observadores.

Cohen desenvolveu essa métrica para a avaliação de dois observadores, também chamados de juízes na literatura. Há duas formas de se calcular o Kappa de Cohen. No primeiro método, o original, não se utiliza pesos em sua análise; enquanto o segundo método utiliza uma ponderação (COHEN, 1968). Ele defendia a ideia de que existem classificações entre os observadores que são mais relevantes do que outras, por isso a necessidade de considerar pesos.

De forma geral o coeficiente Kappa ( $k$ ) sem pesos pode ser calculado a partir de:

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} \quad (1)$$

onde  $\sum fa$  representa o somatório das frequências de concordância observada nas mesmas categorias,  $\sum fe$  representa o somatório das frequências de concordância ao acaso em cada categoria e  $N$  significa o total de eventos observados.

De forma alternativa, o coeficiente Kappa também pode ser obtido como:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

onde  $P_o$  é a proporção de concordância observada, ou seja, a proporção de unidades que os observadores classificaram nas mesmas categorias em relação ao total de classificações, enquanto que  $P_e$  é a proporção de concordância ao acaso, ou seja, a proporção de unidades classificadas pelos observadores nas mesmas categorias por mera coincidência, também em relação ao total de classificações.

Alguns anos mais tarde, FLEISS (1971) estendeu esse comparador para múltiplos observadores, denotando o Kappa generalizado. No entanto, sem pesos. Para este caso, utiliza-se ainda a equação (2), com  $P_o$  e  $P_e$  sendo calculados como:

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i \quad (3)$$

$$P_i = \frac{1}{n(n-1)} [(\sum_{j=1}^t x_{ij}^2) - n] \quad (4)$$

e

$$P_e = \sum_{j=1}^t q_j^2 \quad (5)$$

$$q_j = \frac{1}{nN} \sum_{i=1}^N x_{ij} \quad (6)$$

onde  $N$  significa o número total de objetos observados,  $n$  o número de observadores,  $t$  o número de categorias,  $P_i$  a proporção de concordância de classificações de uma mesma categoria,  $q_j$  é a proporção de todas as atribuições que são de uma determinada categoria  $j$  e  $x_{ij}$  representa o  $i$ ésimo valor na linha  $i$  coluna  $j$  quando os dados estão dispostos na forma de Tabela. Para informações mais detalhadas das equações 3 à 6 sugere-se uma leitura do trabalho realizado por FLEISS (1971).

# CAPÍTULO 3

## ESTADO DA ARTE EM CITOMETRIA DE FLUXO

### III.1 Agrupamento de Dados em Citometria de Fluxo

Ao longo dos anos, diversas técnicas computacionais foram desenvolvidas com o intuito de classificar adequadamente dados de maior dimensionalidade, e em especial, dados de citometria. Seu início deu-se na área biológica, especificamente, na biologia marinha. Esta seção aborda parte dessa trajetória apresentando alguns trabalhos relevantes.

Em 2000, alguns pesquisadores (AL-HADDAD, *et al.*, 2000; MORRIS, *et al.*, 2001) reportaram técnicas de identificação de microalgas marinhas. Geralmente, apenas pequenos conjuntos de dados com diferentes números de observações por grupo taxonômico são possíveis. O objetivo era tentar classificar 20, 40 e 60 espécies de microalgas avaliando a quantidade de dados presentes na etapa de treino. O melhor resultado que os autores conseguiram com redes neurais artificiais foram classificações de 92%, 84% e 77%, respectivamente. Eles concluíram que o aumento no número de neurônios na camada escondida da rede neural artificial teve relativamente pouco efeito na identificação geral e por isso afirmaram que a aplicação de um discriminante linear já seria suficiente.

No ano seguinte, MORRIS e colaboradores (2001) afirmaram que as funções de base radial (*Radial Basis Function* – RFB) são um tipo de rede neural artificial apropriada à tarefa de identificação de dados de grandes dimensões. Para eles, as RBF são fáceis de otimizar e treinam rapidamente. Para identificação, as RBF são hábeis para lidar com novas espécies (BODDY, *et al.*, 1994; WILKINS, *et al.*, 1999). No entanto, quando é

necessário incorporar essas novas espécies, o retreinamento da rede neural consome muito tempo ou é quase impossível de ser feito. O objetivo do mencionado trabalho foi apontar para o fato da técnica de Máquina de Vetores de Suporte (*Support Vector Machines* – SVM) ser mais eficiente que as RBFs para alguns problemas específicos; como, por exemplo, discriminar uma única espécie de fitoplâncton contra um *background* de outras  $N$  espécies. Para isso, eles utilizaram 61 espécies de fitoplânctons com tamanhos variados de 1  $\mu\text{m}$  a 45  $\mu\text{m}$  sob determinadas condições de controle. Para os autores o uso de SVM foi recente, mas pode ser considerada especializada na determinação de limites e fronteiras. Por outro lado, as SVMs ainda não tinham sido bem avaliadas quanto a sua tolerância a ruídos e dados desconhecidos. No estudo em questão, as SVMs propiciaram uma melhor identificação que a RBF (81% contra 77%, respectivamente). Esse estudo analisou dados em 7 dimensões.

Outro trabalho (WILKINS, *et al.*, 2001), também aplicado ao estudo de fitoplânctons, comparou quatro variantes de um algoritmo fuzzy *k-means*. O trabalho identificou problemas comuns com algoritmos iterativos, quando no curso de suas iterações um *cluster* tendia a incorporar outros *clusters*. O algoritmo de DEMERS (DKLL) previne tal situação, obtendo uma taxa de sucesso de 94,8% quando comparado com o padrão ouro. A técnica de Distância Adaptativa (*Adaptive Distance* – AD) também foi uma das variantes deste trabalho e produziu o resultado mais consistente dentre todas as estudadas. A AD conseguiu 96,3% de consistência quando comparado com o padrão ouro. No entanto, quando foi utilizada a técnica da Máxima Verossimilhança como uma variante do algoritmo principal de um *k-means fuzzy*, o resultado piorou para 84,3% quando comparado com o padrão. Uma dificuldade relatada pelos pesquisadores foi quanto à definição no número de *clusters*. Eles se basearam em inspeções subjetivas visuais ou em algum conhecimento a priori dos dados. Assim, os pesquisadores concluíram que métodos para separação de *clusters* não têm sido inteiramente desenvolvidos e que ainda existem dificuldades para se determinar o número ideal de *clusters* para um problema genérico. Este estudo também avaliou dados em 7 dimensões.

Estabelecer um método que estime o número ótimo de *clusters* tem sido um desafio reportado na literatura (WILKINS, *et al.*, 2001; ZENG, *et al.*, 2007; BAI, *et al.*, 2011; PAKHIRA, 2012; LIANG, 2012; YU, *et al.*, 2014). O trabalho de ZENG e colaboradores (2007) estudou a aplicação do algoritmo *k-means* em dados de citometria, reportando que a conformação dos *clusters* resultantes é amplamente dependente do

número inicial de *clusters* ( $k$ ). Para se obter uma estimativa do número apropriado de *clusters* existem duas aproximações:

- Iniciar o algoritmo com um grande número de *clusters* e então diminuí-los com critérios estatísticos, e;
- Iniciar o algoritmo com um pequeno número de *clusters*, mas gradualmente aumentar esse número.

Como já abordado anteriormente, não há um algoritmo de agrupamento de dados universal. É muito importante investigar com cuidado as características da amostra a ser classificada, para selecionar ou planejar uma estratégia de agrupamento de dados apropriada. A diversidade de aplicações reflete a importância do agrupamento de dados na pesquisa científica, causando, por outro lado, certa confusão devido aos seus vários objetivos e terminologias. O algoritmo *k-means*, por exemplo, é baseado em distâncias euclidianas, e por isso tende a gerar *clusters* hiperesféricos; embora *clusters* reais nem sempre possuem essa geometria. A tendenciosidade desta técnica inevitavelmente afeta o desempenho final da classificação. A Figura 5 ilustra o resultado de quatro técnicas computacionais para uma mesma amostra (BODDY, *et al.*, 2001). As imagens estão em escalas arbitrárias e representam a intensidades luminosas das fluorescências de cada canal.

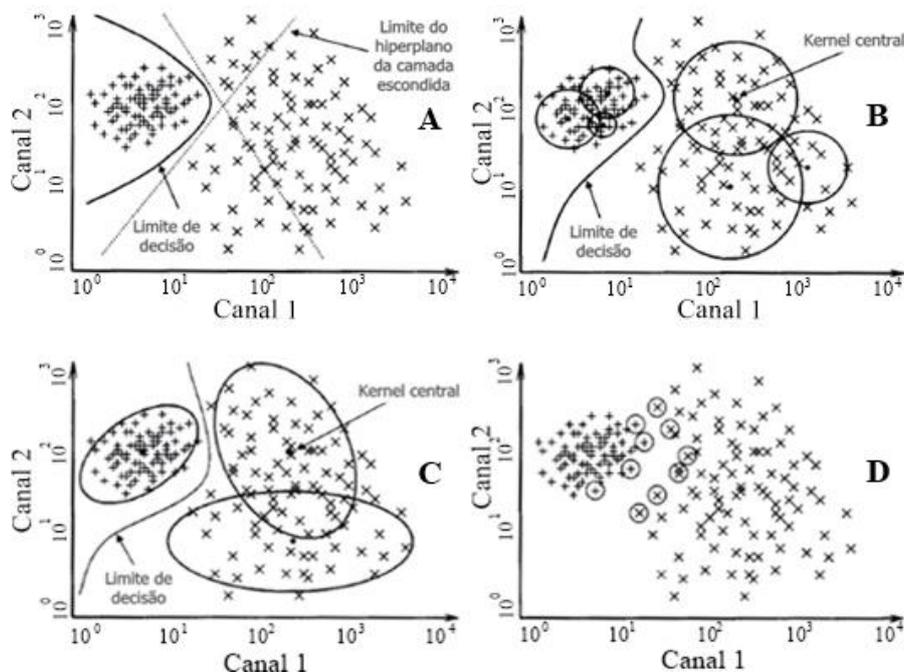


Figura 5. Modos de operação de quatro técnicas computacionais de redes neurais artificiais para uma única identificação de duas dimensões. a) *Perceptron* multicamadas; b) *Radial basis function*; c) *Asymmetric radial basis function*; d) Máquinas de vetores suporte (extraído e adaptado de BODDY, *et al.*, 2001).

No que tange à tentativa de combinação de técnicas computacionais para a solução de algumas limitações específicas, alguns pesquisadores (KITSOS, *et al.*, 2007) propuseram a *Automated High Throughput Platforms e Hierarchical Clustering* para detecção de estágios celulares a partir de dados de citometria. Um dos objetivos do estudo foi avaliar através da técnica de *clusterização* hierárquica a visualização do estado da superfície celular por marcadores celulares. As células estudadas foram de leucemia (*acute promyelocytic leukemia cell line HL-60*). A mineração de dados também permitiu identificar e encontrar perfis de células que não são obviamente detectadas por inspeção visual. A técnica de *clusterização* hierárquica não apenas identifica tendências em grupos de dados, como também revela não linearidades e respostas não óbvias nos dados. Segundo os autores, a habilidade de descobrir perfis celulares atípicos e explorar sua fisiologia relevante é facilitada pela *clusterização* hierárquica e promove um ponto de partida para explorar a biologia de populações de células particulares.

Uma preocupação premente de alguns pesquisadores (LE MEUR, *et al.*, 2007) da área é a qualidade dos dados citométricos. De acordo com os autores, é entendido como qualidade dos dados em citometria de fluxo o cuidado em avaliar como o processo de uso da técnica de citometria de fluxo é realizada. GRATAMA e colaboradores (1998) propuseram algumas diretrizes, tal como monitorar as medidas de fluorescência pelos gráficos de calibração de cada fluorescência. Este entendimento deveria ser adotado como padrão; mas, no entanto, ele não é sistematicamente realizado, ocasionando falta de controle na qualidade desses dados. Ainda segundo LE MEUR e colaboradores (2007), o controle de qualidade e a própria qualidade na avaliação são críticos quando no desenvolvimento de novas tecnologias de alto volume de dados (*High-Throughput – HT*) (KITSOS, *et al.*, 2007). Isto sugere que o gargalo permanece no desenvolvimento de métodos de análise de dados e de visualização em citometria de fluxo de alto volume. A avaliação da qualidade de dados em experimentos de alto volume é complicada exatamente devido a sua alta quantidade de dados. Uma metodologia de avaliação da qualidade é um importante passo para que pesquisadores possam identificar diferenças na amostra original de dados que são provenientes de mudanças do meio e que não é motivada biologicamente. Esse tipo de cuidado estabelece critérios de controle, dando especial consideração a esses dados, ou até mesmo excluindo-os de futuras análises.

LE MEUR e colaboradores (2007) afirmaram também que equipamentos comerciais têm, normalmente, seus próprios programas de aquisição dos dados (por

exemplo: *CellQuest Pro* pela BD *Biosciences*; *Summit* pela *DakoCytomation* ou *Expo32* pela *Beckman Coulter*), sendo essas ferramentas primariamente designadas para um equipamento específico e oferecendo pouca, ou nenhuma, função para se avaliar a qualidade dos dados. A proposta dos autores do mencionado trabalho foi de apresentar métodos gráficos para exploração de dados em 1 ou 2 dimensões. Cinco métodos de visualização distintos foram implementados para explorar a distribuição e densidade de dados de citometria que ainda não tinham tido nenhum tratamento inicial; ou seja, os dados de citometria eram brutos, sem qualquer pré-processamento. O intuito foi demonstrar que ao se analisar dados brutos através dos cinco tipos de gráficos propostos de forma manual, é possível identificar amostras anômalas que não seriam factíveis através de ferramentas automáticas, permitindo ter dados com mais qualidade. Como ressalva, os autores concordaram que esse tipo de processo pode ser difícil e trabalhoso quando a dimensão dos dados é maior do que dois.

LIZARD (2007) relatou que desde a última década numerosas publicações contribuíram para o aumento no conhecimento da citometria de fluxo nas áreas da biologia básica e aplicada, bem como na área médica. No entanto, apenas alguns trabalhos na área de citometria de fluxo foram publicados demonstrando o interesse de usar e desenvolver novas ferramentas analíticas para tratar e interpretar os dados citométricos. O autor ressaltou um artigo da *Cytometry Part A* (2007), “*Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data*”, o qual se preocupou com a análise de técnicas de agrupamento de dados e de componentes principais na área de dados citométricos, enfatizando a importância da obtenção de dados citométricos de alta qualidade. Ele também destacou o interesse de *softwares* comerciais em favor deste tipo de análise, o qual antes somente era realizado por poucos laboratórios especializados. Isto demonstra a preocupação da comunidade científica em querer avaliar melhor dados de citometria. Com o passar dos anos foi possível notar este tipo de preocupação entre os pesquisadores. Como exemplo, GASSEN e colaboradores (2015) avaliaram a qualidade de seu algoritmo comparando com arquivos de referência da primeira chamada de desafios do FlowCAP. O FlowCAP é uma resposta da comunidade científica para testar a qualidade das classificações de algoritmos automáticos em citometria de fluxo. Este evento promove o teste de algoritmos de terceiros em duas etapas: a primeira compara os resultados do algoritmo com análises manuais de especialistas comuns, enquanto que a segunda etapa compara com resultados

clínicos/biológicos. Outros pesquisadores (QIU, 2015) com o mesmo tipo de preocupação participaram de outras chamadas do FlowCAP, demonstrando a continuidade do programa que teve sua primeira iniciativa em 2010 e está em sua quarta chamada atualmente.

Outro tipo de abordagem sobre o estado da arte diz respeito às vantagens obtidas na classificação de dados por algoritmos automáticos não-supervisionados em contraposição aos *gates* manuais. Quando *gates* manuais são utilizados é possível que eventos que não pertençam a um ou outro cluster sejam incorporados em outros clusters de forma equivocada, devido à limitação visual de sua análise em 1 ou 2 dimensões. Quando se utilizam algoritmos multidimensionais essas situações são naturalmente excluídas. No entanto, alguns autores (CHAN, *et al.*, 2008) afirmam que *gates* manuais e algoritmos automáticos multidimensionais não são necessariamente excludentes. Para dados complexos pode ser uma interessante estratégia explorar seus eventos por meio de um especialista para reduzir seu tamanho e sua complexidade com *gates* manuais. Para só então aplicar um algoritmo automático multidimensional. Até porque algoritmos automáticos não-supervisionados não possuem a habilidade de um especialista humano em realizar generalizações (MALEK, *et al.*, 2014). Este tipo de procedimento semiautomático permite carrear o conhecimento do especialista em um estágio inicial da análise dos dados, melhorando o processo de classificação dos dados.

### **III.2 Ferramentas Acadêmicas**

Devido a vários pontos relatados por alguns trabalhos mencionados anteriormente, houve uma natural resposta da academia em solucionar lacunas técnicas em citometria. Além disso, o uso de ferramentas abertas e gratuitas facilita a comparação de resultados biológicos e também a comparação entre diferentes técnicas de agrupamento. Neste contexto podem ser mencionadas três ferramentas: flowClust, FIND e SamSPECTRAL.

O flowClust foi desenvolvido por LO, HAHNE, *et al.*, (2009) com o intuito de suprir a escassez de estatística em citometria e com o foco de construir uma boa ferramenta para seleção automática de *gates*. Ele foi desenvolvido como um pacote de *scripts* do programa R. Os autores continuaram a aplicação utilizando Modelo de Mistura de Gaussianas (GMM – do inglês *Gaussian Mixture Model*) juntamente com o Critério

de Informação Bayesiana (BIC, do inglês *Bayesian Information Criterion*) (SCHWARZ, 1978), para estimar o número inicial de *clusters*. Partes do programa foram desenvolvidas em C para melhorar o desempenho. Para um melhor entendimento sobre o BIC, sugere-se a leitura da secção IV.2.

DABDOUB, RAY e outros estudiosos (2011) desenvolveram o aplicativo FIND, que pode ser também utilizado na plataforma R. O objetivo do desenvolvimento do programa foi ele se tornar uma plataforma de análise para pesquisadores. Sua interface amigável, colorida e de fácil utilização são atrativos para que vários pesquisadores possam utilizar uma ferramenta dedicada à análise de dados de citometria de fluxo. O FIND possui implementado ferramentas do tipo *box-plot*, *scatter-plot* e mapa de calor para gráficos de duas dimensões, por exemplo. No presente trabalho o FIND não foi utilizado dentro do software R, mas sim de forma independente como um arquivo executável.

Por último, o objetivo do desenvolvimento do SamSPECTRAL (ZARE, *et al.*, 2010) foi o de identificar automaticamente populações adjacentes de células, especialmente quando agrupamentos de baixas e altas densidades estão perto uns dos outros. Para os autores, outros programas investigativos apresentam claros problemas em seus métodos de agrupamento porque utilizam técnicas baseadas em modelos, como o FLAME, o flowClust e o flowMerge. A *clusterização* espectral é uma técnica não paramétrica que evita os problemas da estimação de distribuição de probabilidade pelo uso de heurística baseada em grafos. Segundo os autores, esta técnica também não é sensível a *outliers* ou ao formato dos *clusters*. Ela é ajustável e por isso pode ser utilizada em um problema genérico. O parâmetro de escala (*normal.sigma*) e o fator de separação (*separation.factor*) são os dois principais parâmetros que precisam ser ajustados para a procura de células raras. Durante seu desenvolvimento os autores identificaram que o SamSPECTRAL é bem adequado para populações com baixo número de amostras com altas dimensões. No artigo o teste gastou menos de 25 minutos com 100.000 células em 23 dimensões. O computador utilizado tinha um processador de 2.7 GHz. Sua plataforma de operação é o R.

### III.3 Limitações da Análise Clássica

Quando se trabalha na área de imunologia com a técnica de citometria de fluxo no estudo e pesquisa das principais subpopulações das células B, as vistas 2-D mais frequentemente utilizadas para este tipo de análise estão apresentadas na Figura 6A.

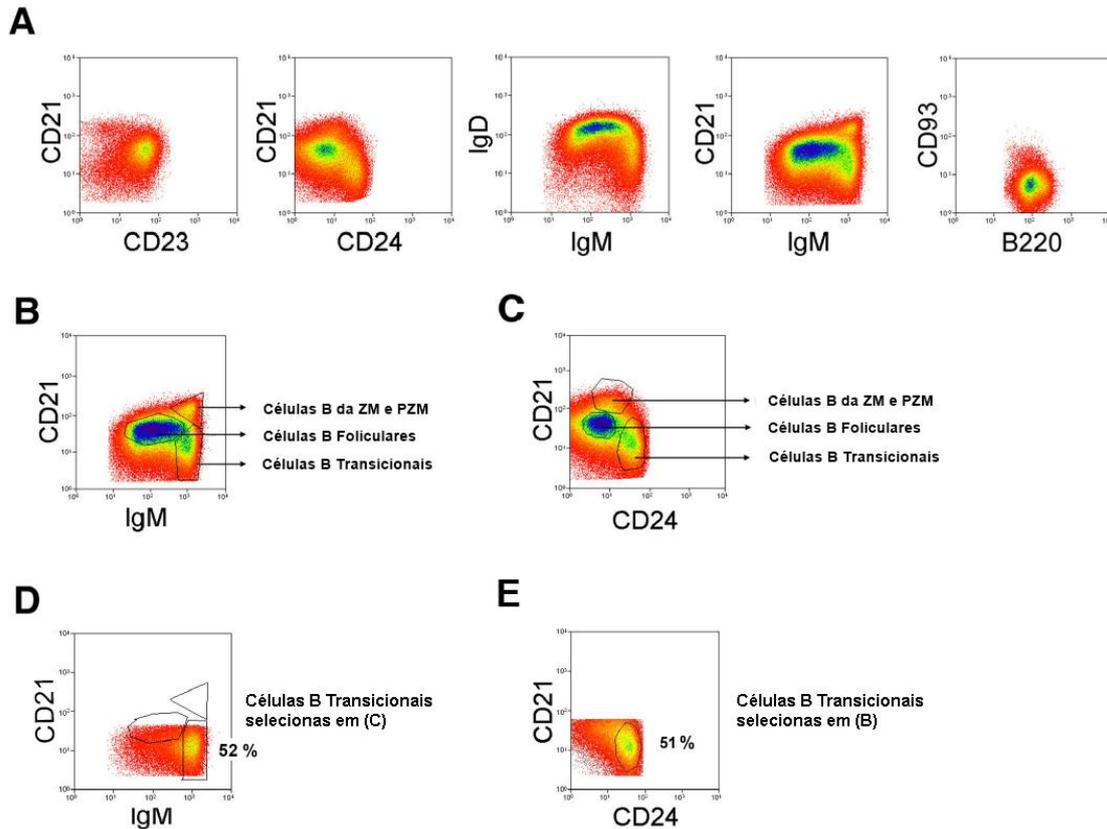


Figura 6. A) Principais vistas 2-D mais utilizadas na imunologia para detecção de células B da zona marginal, transicional e folicular; as imagens B) e C) representam duas abordagens para detecção das células B transicionais. D) apresenta o percentual de células transicionais (52%) que são encontradas na projeção IgM x CD21 pelo *gate* da Figura B) das células B transicionais desta Figura. E) apresenta o percentual de células transicionais (51%) que são encontradas na projeção CD24 x CD21 pelo *gate* da Figura C) das células B transicionais desta Figura.

É importante ressaltar que diferenças entre grupos de pesquisa estão no formato, e onde, esquemas de *gates* são definidos para se classificar os três maiores subconjuntos de células B: folicular, zona marginal e a transicional (TR) (ALLMAN e PILLAI, 2008). Isto leva a algumas incoerências na definição dessas subpopulações, pois dependendo da combinação de *gates*, as classificações das subpopulações não são as mesmas. Por exemplo, as Figuras 6B e 6C apresentam duas abordagens para delimitar as células transicionais (por exemplo). Numa das abordagens caso as células *gateadas* na Figura 6C fossem projetadas na vista IgM x CD21, apenas 52% dessas células seriam consideradas transicionais pelo *gate* da Figura 6B. O restante dessas células se encontraria fora da

região deste *gate* em toda a expressão do IgM. De forma análoga, se as células *gateadas*, na Figura 6B fossem projetadas na vista CD24 x CD21, apenas 51% dessas células seriam consideradas transicionais pelo *gate* da Figura 6E. Isto torna claro as discrepâncias ditas advindas dos diferentes esquemas de *gates*. A seleção de um par de marcadores automaticamente exclui todos os outros. Por isso, contornar essas discrepâncias e tentar conciliar as diferentes formas de visão numa aproximação multidimensional que defina subpopulações de *clusters* levando em consideração todos os marcadores simultaneamente é algo bastante desejável.

A falta de padrão nos procedimentos para classificações é uma preocupação da área (HASSAN, *et al.*, 2015). Este tema foi o assunto mais discutido no I Workshop de Citometria de Fluxo do EUROFLOW, realizado em agosto de 2011 no Rio de Janeiro (Brasil) e que contou principalmente com a presença de pesquisadores brasileiros e europeus.

Análises baseadas em vistas 2-D puramente canônicas impõem outra limitação que frequentemente não é comentada. A projeção canônica de dados em 2 dimensões pode não revelar a existência de dados agrupáveis que somente podem ser separáveis em dimensões acima do 2-D. Sua dificuldade de visualização pode ser parcial, ou até mesmo total. Esta limitação pode ser particularmente crítica para a identificação de células menos numerosas, tais como a pré-zona marginal de células B, ou as células B transicionais posteriores, por exemplo. Uma abordagem multidimensional onde todos os parâmetros são analisados simultaneamente pode obter uma aproximação mais fidedigna da realidade.

# CAPÍTULO 4

## ALGORITMOS ADAPTADOS

O objetivo deste capítulo é descrever os métodos de agrupamento de dados desenvolvidos ou utilizados no presente trabalho.

### IV.1 Inicialização de Centroides Baseada em Modas

LO e colaboradores (2008) reportaram que agrupamentos de dados baseados em modelos produzem bons resultados para dados de grande dimensionalidade (tipicamente maiores que 10). E, segundo eles, modelos baseados em Mistura de Gaussianas são boas referências para este tipo de agrupamento. Entretanto, para esta técnica, ainda existem dois fatores bastante graves. O primeiro refere-se ao número de *outliers*, que podem tendenciar o resultado final. O segundo fator se refere ao número inicial de *clusters* que o algoritmo deve considerar em seu estágio inicial. Uma alternativa interessante é iniciar o algoritmo com um número baixo de *clusters* e ir aumentando gradativamente este número até o limite máximo de *clusters* possíveis. Então, a cada aumento de um *cluster* no algoritmo guardam-se os resultados da iteração atual. Ao final do programa, escolhe-se a iteração com o melhor resultado, o qual pode ser estimado utilizando uma função custo pré-determinada. Esta alternativa, além de ser morosa, pode ser ineficiente por não se saber boas posições iniciais dos centroides. Como citado neste trabalho, os autores informaram que uma má inicialização dos centroides aumenta o tempo de convergência do algoritmo ou o faz cair em um mínimo local. Eles utilizaram o BIC (SCHWARZ, 1978) para estimar o número inicial de *clusters*, assim como FRELINGER *et al.* (2010) em trabalhos anteriores. Utilizando um PC com processador Intel de 3 GHz e 2 GB RAM, eles levaram 20 minutos para analisar uma amostra de 1 milhão de células com 10 dimensões. A tentativa de estimar o número ideal de *clusters* antes de um agrupamento

de dados é uma medida importante para facilitar o trabalho do algoritmo em encontrar a melhor classificação possível.

O presente trabalho seguiu filosofia semelhante, com a diferença de em vez de ser sugerido um número inicial de *clusters*, o algoritmo de identificação de modas (descrito mais à frente) procurou informar o número máximo de *clusters* possíveis em um determinado conjunto de dados. A partir daí o usuário pode escolher o número ideal de *clusters* que ele supõe ser adequado. Esta forma de abordagem traz a vantagem adicional de poder adicionar a *expertise* do usuário numa decisão inicial. Essa forma de agir traz à tona uma vertente deste trabalho, que é a de ter uma ferramenta de agrupamento de dados semiautomática, onde o conhecimento intrínseco e acumulado do usuário é trazido para a ferramenta com o intuito de melhorar o processo de agrupamento.

O princípio acima mencionado norteou o método desenvolvido para a determinação do número de *clusters* e a posição inicial de seus centroides.

Tomemos como exemplo, duas amostras hipotéticas de dados bidimensionais, denominadas amostras A e B, representadas pelos gráficos de espalhamento ilustrados na Figura 7. As unidades da Figura são arbitrárias.

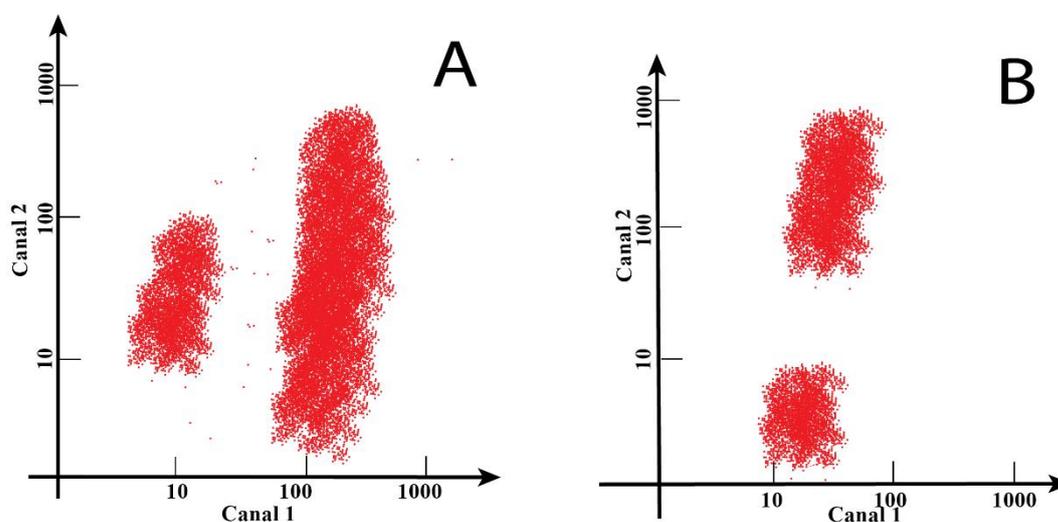


Figura 7. Gráficos de espalhamento de duas amostras de dados. A Figura do lado esquerdo é associada à amostra A e a Figura do lado direito à amostra B. Os eixos x e y são arbitrários.

É fácil perceber que em ambas as amostras existem dois *clusters* bem distintos. A pergunta que se coloca é: *como desenvolver um algoritmo automático que detecte a quantidade máxima de clusters, em qualquer dimensão, informando ao mesmo tempo a*

*melhor localização inicial de seus centroides?* Caso isso seja possível, o agrupamento final obtido por uma técnica qualquer seria melhorado por partir de dados iniciais mais adequados (BARBARA, 2009). Além disso, como já teria sido estimada uma boa posição inicial dos centroides, o tempo de execução dessa classificação tornar-se-ia menor (PICARD, 2007).

O início do algoritmo de determinação do número de *clusters* e de inicialização de centroides proposto passa pelo cálculo das componentes principais, apontadas pelos autovetores, da matriz de dados ( $\mathbf{v1}$  e  $\mathbf{v2}$ , no caso do exemplo) (Figura 8A). Observando-se a amostra A e projetando-se seus dados sobre cada componente principal nota-se que a amostra A é bimodal na direção apontada por  $\mathbf{v2}$  e monomodal na direção apontada por  $\mathbf{v1}$  (Figura 8B). As unidades da Figura são arbitrárias.

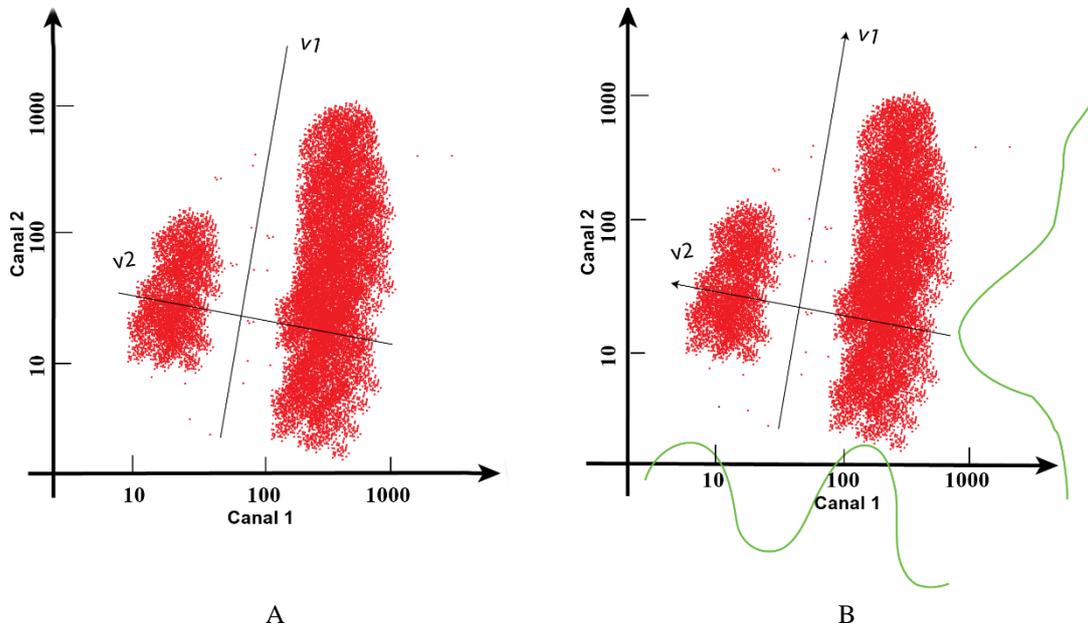


Figura 8. Do lado esquerdo a amostra A com a representação dos seus autovetores  $\mathbf{v1}$  e  $\mathbf{v2}$ . Do lado direito a mesma amostra com a projeção de seus pontos em seus autovetores. Os eixos x e y são arbitrários.

Observando-se a amostra B nota-se que ela tem um comportamento similar (Figura 9A). No entanto, ela se apresenta bimodal na direção apontada por  $\mathbf{v1}$  e monomodal na direção de  $\mathbf{v2}$  (Figura 9B). As unidades da Figura são arbitrárias.

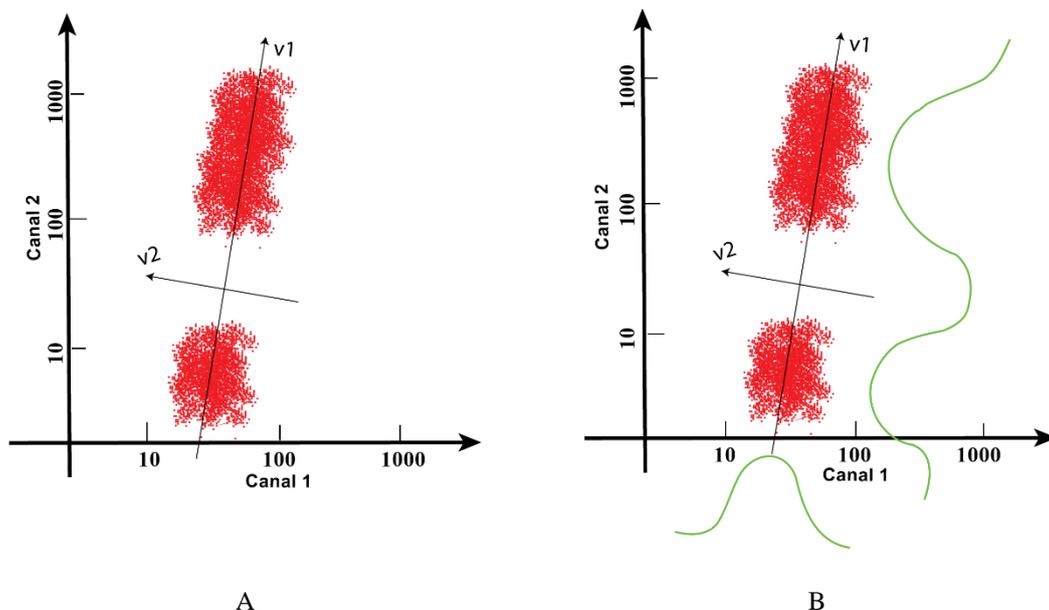


Figura 9. Do lado esquerdo a amostra B com a representação dos seus autovetores  $\mathbf{v1}$  e  $\mathbf{v2}$ . Do lado direito a mesma amostra com a projeção de seus pontos em seus autovetores. Os eixos x e y são arbitrários.

A ideia do algoritmo de inicialização desenvolvido é estimar a quantidade e a posição dos centroides dos *clusters* pela quantidade de modas resultantes da projeção dos dados nas direções das componentes principais. A multiplicação da quantidade de modas em cada componente principal (também relacionado com a quantidade de dimensões da amostra de dados) indicará a quantidade máxima de *clusters* possivelmente existentes na amostra dos dados. O desafio agora é como encontrar automaticamente essas modas a partir dos dados.

Uma vez que os dados estejam projetados na direção de uma componente principal, o algoritmo de inicialização possui uma rotina interna que é responsável por criar um histograma para cada dimensão. Para isso, foi utilizado o estimador de densidade para dados unidimensionais desenvolvido por BOTEV (2010). Ele cria um histograma informando numérica e graficamente como se distribui a frequência dos dados. A Figura 10 ilustra uma situação onde existem três modas representadas pelos seus respectivos picos de frequência. Os valores da abscissa e da ordenada foram suprimidos para efeito didático.

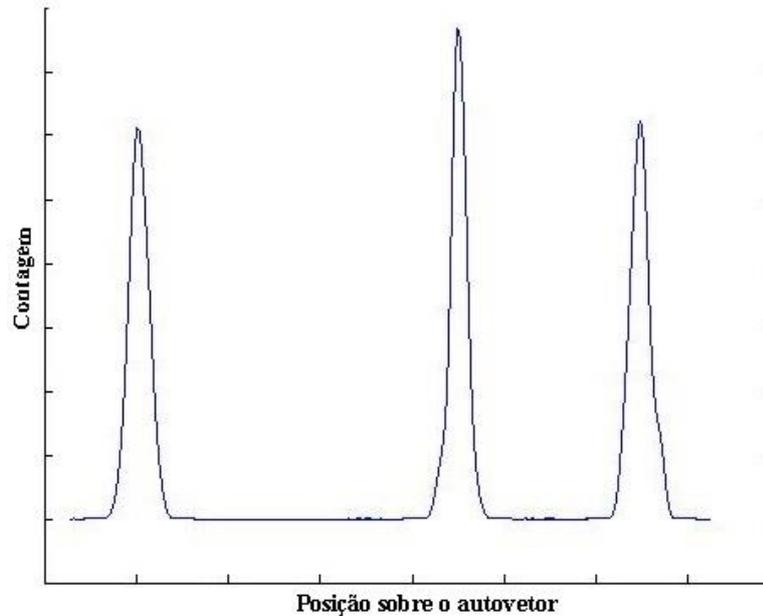


Figura 10. Histograma representativo de 3 modas. Os eixos x e y são arbitrários.

A partir deste histograma facilmente se obtém o valor da 1ª moda, ou seja, a mais expressiva. Isto pode ser feito usando-se, por exemplo, a função `MODE`, do Matlab, que retorna o valor de sua frequência e sua posição.

O próximo passo é calcular a derivada de cada ponto do histograma para obter pontos de inflexão, isto é, onde a concavidade do histograma muda. Este artifício irá sinalizar onde existem outros pontos de máximo, tornando-se novas modas. Na primeira iteração da rotina é guardada a posição da 1ª moda, ou seja, aquela de maior frequência. A partir deste ponto a rotina busca encontrar uma nova moda para um dos lados da 1ª moda. Um novo valor é registrado quando há mudança do sinal da derivada do sinal positivo para negativo na distribuição de frequência. A Figura 11 apresenta essa situação na busca da segunda moda. Quando isso acontece, o valor do pico e a posição da moda são registrados.

Como medida de controle há um limite mínimo de frequência para que um pico seja considerado uma moda, eliminando possíveis ruídos. Caso a altura do pico seja menor do que 5% (cinco por cento) da altura do maior pico, a candidata à moda é rejeitada.

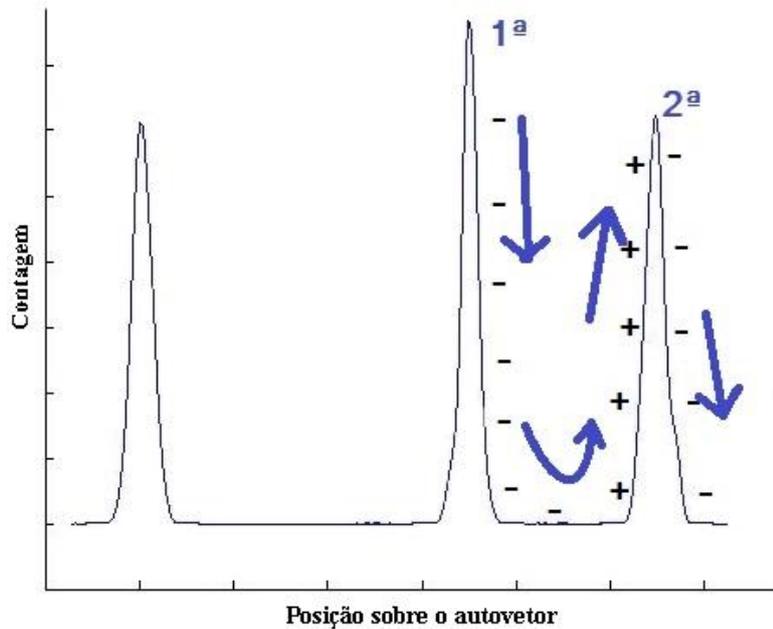


Figura 11. Esquemático de varredura de picos de frequência para o lado direito a partir da moda principal. Os eixos x e y são arbitrários.

Quando a rotina chega à posição limite, seja máxima ou mínima, ela retorna à posição da 1ª moda e executa a segunda (e última) rodada. O procedimento se repete até que se chegue novamente ao limite inferior do histograma, conforme pode ser visto na Figura 12. Neste caso é registrada a 3ª moda para o exemplo ilustrado.

A última etapa dessa rotina é a conversão das posições das modas para os eixos cartesianos reais das amostras, uma vez que as mesmas estão referenciadas nas suas componentes principais.

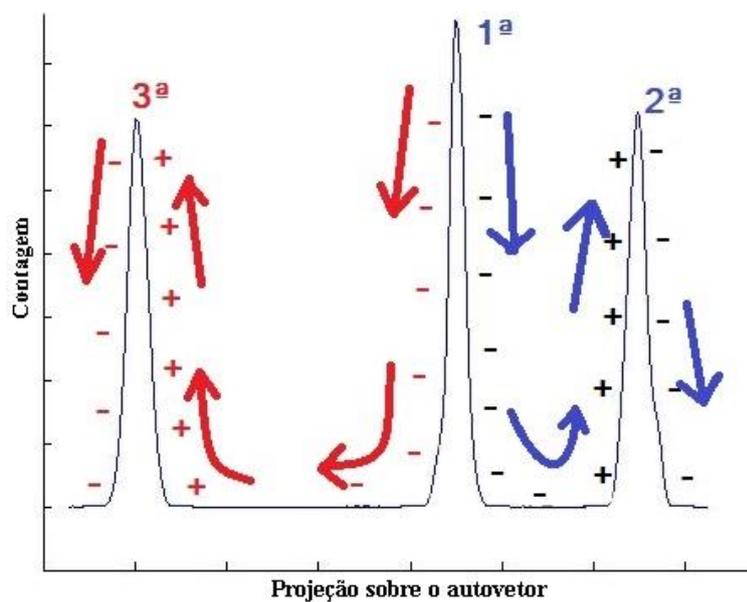


Figura 12. Esquemático de varredura de picos de frequência para o lado esquerdo a partir da moda principal. Os eixos x e y são arbitrários.

O próximo passo do algoritmo de determinação do número de *clusters* e inicialização de centroides é obter todas as possíveis posições de centroides combinando-se as modas de cada dimensão. Isso gera um número total de possíveis *clusters* que é a multiplicação direta do número de modas de cada dimensão. Por fim, estes são ordenados por ordem decrescente dos produtos das frequências de ocorrência das modas. Ou seja, quanto maior o valor da multiplicação na etapa anterior, maior a chance de existir um centroide bem definido na posição combinada de cada moda. Isto significa que nas duas dimensões analisadas há uma região comum com maior concentração de pontos e que por isso, deve haver um *cluster* nesta região com seu centroide inicializado na interseção das projeções das modas. Ao se realizar essa multiplicação todos os possíveis centroides são ranqueados. Este *ranking* é informado ao usuário para que ele possa decidir a quantidade desejada de *clusters* com o algoritmo de agrupamento de dados que ele irá efetivamente operar. O algoritmo associa o número de *clusters* indicado pelo usuário aos centroides mais bem ranqueados, até o limite máximo anteriormente informado.

Apesar do exemplo utilizado para a explanação do método ter sido com apenas duas dimensões, a rotina pode ser utilizada em qualquer dimensionalidade.

#### IV.1.1 Pseudocódigo da Inicialização dos Centroides

1. Início;
2. Identifica quais canais (dimensionalidade –  $k$ ) serão considerados na análise;
3. Cria matriz de autovetores dos dados ( $V$ ) com os canais selecionados e projeta os dados na direção desses autovetores;

$$v_k = (x - \mu) * V_k \quad (7)$$

onde  $v_k$  é projeção dos dados na direção de cada autovetor,  $\mu$  representa o vetor de média  $x$  e  $V_k$  é o  $k$ -ésimo autovetor;

4. Chama subrotina de BOTEV (2010) para criar histogramas de frequências;
5. Obtém 1ª moda

$$modas(1) = mode(v_d)$$

6. Calcula a derivada de todos os pontos do histograma e registra picos de máximo

do

$$der_k = derivada(v_k)$$

end

for  $i=1:n$

```

        if derk(i)= sinal positivo
            if derk(i+1)= sinal negativo
                [posi(i) freq(i)] = vk(i) %registra posição e valor do pico
            end
        end
    end
end

```

7. Elimina picos muito pequenos que podem ser ruídos graças a uma variável chamada *fator\_de\_corte*. Geralmente essa variável tem valor 0,05. Ela é referenciada à primeira moda por esta ser a maior.

```

    for i = 1:length(freq)
        if freq(i) > fator_de_corte*moda(1)
            freq(i) = freq(i) % mantém valor
        else
            freq(i) = []; % exclui valor
        end
    end
end
modas = (posi , freq)

```

8. Combina todos os picos de frequência de cada canal para estabelecer parâmetro de maiores pontos de frequência coordenados. Para apenas 2 canais, temos;

```

temporario = modas(freq(1)).*modas(freq(2))

```

9. Ordena as modas e define regiões mais prováveis de centroides iniciais;

```

modas = ordena(temporario)

```

10. Translada as modas para o eixo real;

```

centroides_possiveis = Vk*modas + μ

```

11. Apresenta a quantidade máxima de possíveis centroides e pergunta para o usuário a quantidade de centroides (*t*) que ele estima realmente ter

```

centroides_iniciais = centroides_possiveis (de 1 até t)

```

12. Fim

O algoritmo desenvolvido propõe automaticamente um número máximo de centroides, que foram estimados pelas modas nas componentes principais de seus dados. Este número máximo é o limite superior de rodadas do BIC, eliminando assim um parâmetro subjetivo para o usuário. Além disso, as posições iniciais fixas desses centroides fornecem apenas uma única tentativa da técnica de BIC em cada dimensionalidade, tornando-se, portanto, um algoritmo determinístico.

## IV.2 Bayesian Information Criterion (BIC)

O BIC é uma métrica baseada em modelos que escolhe o menor valor de sua função custo (Ramírez, 2006).

$$\text{BIC} = -\log P(D|\Theta) + d/2 * \log n \quad (8)$$

onde  $P(D|\Theta)$  significa a função de verossimilhança,  $D$  significa os dados,  $\Theta$  e  $d$  a significam os parâmetros e a dimensão do modelo, respectivamente e  $n$  representa o tamanho da amostra.

O valor de BIC considera dois aspectos bastante interessantes, acurácia e complexidade. A primeira parte da equação 8 refere-se à acurácia e depende do quanto o modelo se adequa aos dados. A segunda parte refere-se à complexidade, penalizando o valor do BIC se a dimensão do modelo aumenta.

Para se obter o número ótimo de *clusters*, normalmente os algoritmos utilizam a técnica de BIC executando vários agrupamentos com o mesmo número de centroides variando apenas suas posições iniciais a cada novo teste. Um grupo de testes é chamado de rodada e ao final da rodada é salvo o melhor (menor) valor do BIC para esta determinada quantidade de centroides. O modelo é então incrementado com mais centroides e novas rodadas são realizadas até um número máximo de dimensões do modelo. Finalmente cada modelo apresenta seu menor valor de BIC. O menor valor de BIC entre todos os modelos representa o número ótimo de *clusters*.

O procedimento acima descrito tem duas grandes considerações. A primeira é relacionada a inicialização aleatória de cada centroide em cada etapa do algoritmo que procura o melhor valor de BIC para uma dada dimensão do modelo. De acordo com alguns autores (ARAI e BARAKBAH, 2007), centroides gerados aleatoriamente não garantem um ótimo global, mas apenas mínimos locais (KOVESI, BOUCHER e SAOUDI, 2001). A segunda consideração é relacionada ao tempo gasto para processamento de toda a técnica. LO *et al.* (LO, BRINKMAN e GOTTARDO, 2008) apontaram que uma má inicialização dos centroides aumenta o tempo de convergência do algoritmo. Adicionalmente, não é claro a definição sobre quantas vezes devem ser testadas as posições iniciais dos centroides afim de se garantir o melhor resultado possível.

### IV.3 Coeficiente Individual Kappa (CIK)

Um objetivo importante deste trabalho é apresentar uma metodologia para se agrupar dados de forma quasi-determinística utilizando o coeficiente Kappa. No presente trabalho a expressão quasi-determinística é usada para expressar uma técnica de agrupamento de dados que produza resultados de agrupamento iguais (determinísticos) mesmo quando a inicialização de seus centroides tenham sido feita de modo aleatório. Este tipo de metodologia se torna principalmente interessante quando não há um método determinístico para a estimação das posições iniciais dos centroides, obtendo-se vários agrupamentos potencialmente equivalentes entre si. É sabido que diferentes inicializações dos centroides levam a etapas de agrupamento de dados diferentes (BARBARA, 2009), pois basta que em uma rodada um, ou mais, centroides tomem uma direção diferente de outra rodada qualquer, para que todo o seu percurso possa ser alterado até a finalização do processo de agrupamento.

A utilização do coeficiente Kappa permite que se meça o grau de concordância entre diversos observadores atribuindo um valor de concordância geral. Apesar da literatura apontar duas faixas para valores Kappa, indo de -1 a 1, ou de 0 a 1, (POWERS, 2012; SIM e WRIGHT, 2005; McHUGH, 2012) este trabalho irá optar apenas pela segunda representação (de 0 a 1) por questões práticas. No entanto, este coeficiente Kappa é um indicador global entre observadores, e não um indicador individual de cada evento de uma amostra. Com o intuito de deixar claro, relembremos a equação (4) que descreve que:

$$P_i = \frac{1}{n(n-1)} [(\sum_{j=1}^t x_{ij}^2) - n] \quad (9)$$

onde  $P_i$  é a variável que calcula o grau de concordância de cada evento de uma amostra. Sua relevância neste contexto é que a partir dele é possível medir quanto cada observador concorda (ou não) com a classificação de cada evento da amostra, realizando, portanto, uma análise individualizada. Vale lembrar que este indicador também varia entre 0 e 1.

A utilização deste indicador ( $P_i$ ) torna-se bastante útil porque a geração aleatória dos centroides iniciais não garante que os agrupamentos finais sejam sempre os mesmos. No entanto, estabelecer um valor para  $P_i$  faz com que apenas os dados que tenham esse valor mínimo possam ser considerados como dados válidos. Neste trabalho cada dado

representa uma célula e chamaremos  $P_i$ , a partir daqui, de Coeficiente Individual Kappa (CIK).

Este tipo de abordagem torna-se interessante quando se deseja diminuir a dúvida sobre regiões de fronteira entre grupos. Assim como é fácil entender que células mais próximas aos centroides possuem uma alta probabilidade de pertencerem ao centroide que os representa, não é difícil compreender que quanto mais afastadas essas células estão, se aproximando de outros centroides, tendem a ter uma maior probabilidade de não pertencerem ao seu agrupamento original. Por isso, regiões limítrofes entre *clusters* tendem a ter células com maior variabilidade entre grupos. Esta abordagem busca minimizar dados que sejam dúbios, selecionando apenas os que tenham um CIK pré-determinado pelo usuário. Isto proporciona uma maneira elegante de se trabalhar com centroides iniciais aleatórios sem que os agrupamentos finais sejam tão aleatórios devida a este tipo de inicialização. Sua desvantagem está em ter de especificar uma quantidade dita como suficiente para minimizar o efeito da aleatoriedade das posições iniciais desses centroides. Esta ação aumenta o tempo de processamento final. Análises conservadoras tenderão a escolher um CIK próximo ou igual a 1, enquanto análises mais tolerantes poderão optar por valores próximos de zero.

Ainda que esta linha de pensamento atribua um CIK a cada célula biológica (dado) que possa ser entendido até como um valor percentual de probabilidade, esta metodologia não pode ser considerada uma técnica de Lógica Fuzzy clássica (ZADEH, 1965). O “Kappa modificado” não informa qual o grau de probabilidade de uma célula biológica pertencer a um outro grupo; mas sim, o grau de concordância de uma célula em pertencer a um *cluster* de acordo com a “opinião” dos atuais observadores. Mudando-se os observadores, a concordância pode mudar.

Dizemos que este tipo de classificação é quase-determinística porque não é possível estimar a quantidade mínima de classificações que garanta um comportamento determinístico. Sempre será possível ter uma nova classificação (novo observador) com células classificadas diferentemente, diminuindo o CIK dessas células.

#### **IV.4 Modelo de Mistura de Gaussianas (MMG)**

Ambos os métodos descritos neste capítulo utilizaram o Modelo de Mistura de Gaussianas (MMG) para modelagem da distribuição dos dados no espaço de parâmetros

e para a etapa de classificação e agrupamento. Aqui será detalhada esta técnica de modelagem e como ela foi incorporada ao programa de computador desenvolvido neste trabalho.

#### IV.4.1 Distribuição Gaussiana

A distribuição Gaussiana, ou distribuição normal, é amplamente usada como modelo de distribuição de variáveis contínuas (BISHOP, 2006). Para uma matriz  $\mathbf{A}$  de dados vetoriais de dimensão  $D$ , a distribuição tem a seguinte forma,

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (10)$$

onde  $\boldsymbol{\mu}$  é vetor de média de dimensão  $D$ ,  $\boldsymbol{\Sigma}$  é a matriz de covariância ( $D \times D$ ) da matriz de dados  $\mathbf{A}$  e  $|\boldsymbol{\Sigma}|$  é o determinante de  $\boldsymbol{\Sigma}$ .

A distribuição Gaussiana é amplamente utilizada como um modelo de densidade e ela possui alguns aspectos importantes. Como primeiro ponto, vamos considerar o número de parâmetros livres da distribuição. Uma matriz de covariância simétrica geral terá  $D(D + 1)/2$  parâmetros independentes e existirão outros  $D$  parâmetros independentes da média ( $\boldsymbol{\mu}$ ), totalizando  $D(D + 3)/2$  parâmetros para caracterizar a distribuição. Para uma amostra de alta dimensionalidade, o número total de parâmetros cresce quadraticamente com  $D$  e com isso o custo computacional para realizar o cálculo das matrizes inversas torna-se proibitivo (BISHOP, 2006) quando este é necessário várias vezes. Restringindo-se a forma da matriz de covariância é possível contornar essa limitação. Se considerarmos que as matrizes de covariância são do tipo diagonal,  $\boldsymbol{\Sigma} = \text{diag}(\delta_i^2)$ , teremos um total de parâmetros independentes igual a  $2 \cdot D$  ( $D$  médias e  $D$  variâncias). Para o caso bidimensional mostrado na Figura 13B, isso corresponde graficamente ao contorno da distribuição ser alinhada aos dois eixos coordenados. Uma restrição ainda maior seria considerar a matriz de covariância apenas proporcional à matriz identidade,  $\boldsymbol{\Sigma} = (\delta^2 \mathbf{I})$ , conhecida como covariância isotrópica, obtendo-se agora um número total de parâmetros independentes que é igual a  $D + 1$  ( $D$  médias e 1 variância). Essa nova aproximação simplifica bastante o custo computacional. Essa distribuição terá um comportamento de superfícies esféricas (Figura 13C). Todas essas três possibilidades podem ser entendidas graficamente pela Figura 13.

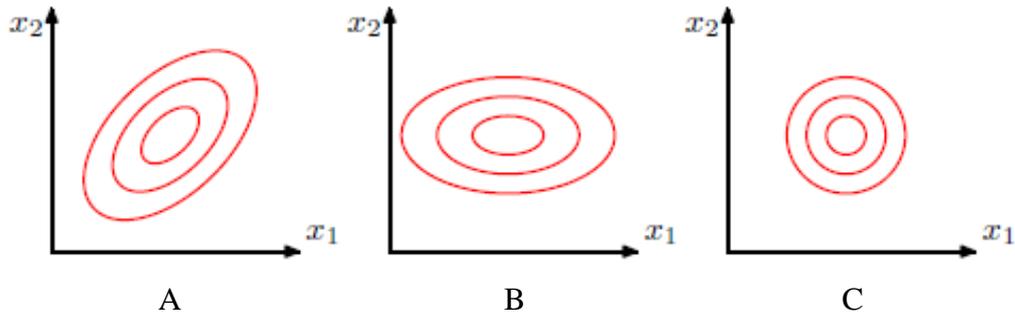


Figura 13. Contorno da distribuição Gaussianas em duas dimensões nas quais a matriz de covariância é A) geral, B) diagonal (alinhado ao eixo das coordenadas), C) proporcional a matriz identidade (extraído de BOTEV, 2006). Os eixos  $x$  e  $y$  são arbitrários.

Apesar das aproximações acima cercearem a forma da densidade de probabilidade da amostra analisada e de limitarem a conformação de *clusters* de dados reais, elas proporcionam, ao menos, formas mais rápidas de se realizar o cálculo da inversa da matriz de covariância. Dependendo do objetivo este tipo de simplificação pode ser bastante interessante.

#### IV.4.2 A Mistura de Gaussianas

O Modelo de Mistura de Gaussianas (MMG) é um modelo estocástico que modela a distribuição de dados multidimensionais de maneira muito sofisticada, sendo capaz de definir distribuições complexas (VEZHNEVETS, *et al.*, 2003). Ele é amplamente utilizada em mineração de dados, reconhecimento de padrões, aprendizagem de máquina e em análise estatística (BISHOP, 2006). Segundo MARTINEZ e MARTINEZ (2008) e BARBARA (2009) uma das principais vantagens do MMG é seu pouco custo computacional quando utilizado com dados multidimensionais. Ainda segundo BISHOP (2006), em muitas aplicações seus parâmetros são determinados pelo método da máxima verossimilhança, tipicamente utilizando o algoritmo da *Expectation-Maximization* (EM), o qual será melhor descrito adiante.

Quando o objetivo é trabalhar com uma mistura de gaussianas, alguns parâmetros adicionais devem ser informados, como a quantidade de gaussianas presentes ( $K$ ) e o peso de cada gaussianiana ( $w$ ). Esta formulação pode ser entendida por:

$$\mathbf{p}(\mathbf{x}) = \sum_{k=1}^K \omega_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (11)$$

onde

$$\sum_{k=1}^K \omega_k = 1 \quad (12)$$

Cada densidade da Gaussiana  $N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$  é uma componente da mistura e tem seus próprios parâmetros de média e matriz de covariância. No entanto, os valores verdadeiros dos parâmetros são desconhecidos, a priori.

#### IV.4.3 Algoritmo *Expectation-Maximization* (EM) para Gaussianas

Como anteriormente mencionado, o algoritmo EM é o método mais utilizado para a realização de estimação de parâmetros de MMGs, tendo sido descoberto e empregado independentemente por diferentes pesquisadores, até DEMPSTER e colaboradores (1977) desenvolverem suas ideias em conjunto, provando a convergência e criando o termo: algoritmo de EM (MOON, 1996). De acordo com BARBARA (2009), é necessário iniciar os parâmetros do algoritmo de EM com valores adequados, uma vez que a convergência do algoritmo depende desses valores iniciais. PICARD (2007) também afirmou que essa técnica é sensível aos valores iniciais, além de informar que a taxa de convergência pode ser lenta. Mesmo assim, o EM é um método para encontrar soluções por máxima verossimilhança de variáveis latentes (DEMPSTER, *et al.*, 1977; McLACHLAN e KRISHNAN, 1997).

Este algoritmo é dividido em dois passos. O primeiro passo é chamado de Esperança (*E*) e é nessa etapa que o algoritmo utiliza os parâmetros atuais das  $k$  gaussianas da MMG e as observações (os dados) para calcular probabilidades a posteriori (BISHOP, 2006). Já o segundo passo é chamado de Maximização (*M*), onde são aplicadas as probabilidades da etapa anterior para maximizar os valores mais prováveis das estimativas dos parâmetros (GEJGUS, *et al.*, 2004). Esses dois passos agem de forma iterativa até que a convergência seja alcançada.

Para o algoritmo EM a convergência pode ser entendida como: a cada iteração do algoritmo, os valores dos parâmetros são computados de modo que a função de verossimilhança sempre incremente seu valor (MOON, 1996). Isso significa que a cada iteração os parâmetros estimados fornecem um incremento da função de verossimilhança, até um máximo local ser alcançado. Não há garantias de que a convergência será um máximo global. Para a função de verossimilhança com múltiplos máximos, a convergência para um máximo local depende dos parâmetros iniciais. Para uma mistura de gaussianas aplicadas ao problema de agrupamento de dados, os parâmetros iniciais são sua quantidade de gaussianas e os vetores de média (posição de um centroide em um

*cluster*). Além da matriz de covariância (informação relacionada com o espalhamento, ou dispersão, do *cluster*) que é calculada pelo algoritmo de EM.

A quantidade  $k$  de componentes de gaussianas do MMG ainda é um dado desconhecido a priori. Sendo assim, com o intuito de se utilizar, de forma mais eficiente, o Modelo de Mistura de Gaussianas com o algoritmo EM é que desenvolvemos o método não supervisionado baseado em modas para estimar estatisticamente o número de gaussianas iniciais, bem como a posição inicial de cada centroide apenas a partir dos dados apresentados.

#### **IV.4.4 Classificação dos dados a partir do MMG**

Após a aplicação do algoritmo de EM são criados tantos modelos quantos necessários segundo a equação 10. Para cada um dos modelos se calcula a probabilidade do dado  $x$  ser descrito por cada um dos modelos. Finalmente, a classe atribuída ao dado será aquela que correspondente ao modelo que apresentar a maior probabilidade. A equação (13) descreve o procedimento utilizado para realizar a classificação dos dados de acordos com  $K$  modelos MMG.

$$c(x) = \max \sum_{k=1}^K p_k(x) \quad (13)$$

onde,

$c(x)$  representa a classe atribuída a cada dado,  $k$  representa o modelo,  $K$  o número total de modelos a serem analisados e  $p_k(x)$  significa a probabilidade do dado ser explicado pelo  $k$ -ésimo modelo.

# CAPÍTULO 5

## MATERIAIS E MÉTODOS

A literatura aponta (PANGBORN, 2009; BISHOP, 2006) que não existe uma única técnica, ou algoritmo, para agrupamento universal de dados. A melhor orientação tem sido partir-se das características de um problema e encontrar a técnica computacional mais adequada, buscando-se uma solução particular para o problema. Por isso, buscou-se desenvolver técnicas que pudessem contribuir para a melhoria de métodos de agrupamento de dados multidimensionais e, em especial, de citometria de fluxo.

As técnicas desenvolvidas resultaram em dois métodos multidimensionais, não supervisionados, ora automáticos e ora semiautomáticos, e que se diferem fundamentalmente na forma como os seus centroides foram inicializados e na técnica de análise de dados, dependendo de cada aplicação. Dito isto, podemos enumerá-los como:

- Método de agrupamento de dados com inicialização determinística de centroides baseada em modas;
- Método de agrupamento de dados quasi-determinístico baseado no coeficiente individual Kappa (CIK).

### **V.1 Método de agrupamento de dados com inicialização de centroides determinísticos baseado em modas**

O referido método diz respeito à junção da técnica de inicialização de centroides baseada em modas sugerida no presente trabalho (Seção IV.1), com o emprego do Critério de Informação Bayesiana – BIC (Seção IV.2) para estimação do melhor número de centroides e o método de modelagem e classificação de dados MMG (Seção IV.4).

De modo a representar diferentes graus de dificuldade para se estimar de forma automática o melhor número de *clusters* e também a localização de cada centroide, quatro

conjuntos de dados foram utilizados para avaliar um dado método. Os resultados obtidos foram comparados com aqueles conseguidos com o método clássico de se inicializar aleatoriamente as posições dos centroides em torno do vetor médio dos dados. Três destes conjuntos são de dados simulados e normalmente utilizados para testes de classificação. O último conjunto representa uma amostra real de dados de citometria de fluxo, uma vez que a aplicação pretendida é agrupar dados deste tipo.

Os conjuntos de dados simulados foram obtidos a partir do sítio eletrônico da Universidade Eastern Finland (Computing, 2012) e são chamados de: R15 (Veenman, Reinders e Backer, 2002), *Compound* (Zahn, 1971) e *Aggregation* (Gionis, H. Mannila e Tsaparas, 2007). Todos eles estão em formato de texto (.txt).

O conjunto de dados de citometria de fluxo foi fornecido pela Unidade de Citometria do Departamento de Imunologia da Universidade Federal do Rio de Janeiro (UFRJ), Brasil. Este conjunto refere-se a uma amostra de sangue humano analisada por um citômetro de 7 canais (dimensões), em que 5 destes 7 canais são canais de fluorescência, nomeados de: FL1, FL2, FL6, FL7 e FL8. Os outros dois canais são o espalhamento frontal e lateral, chamado de FSC e SSC, respectivamente. A partir dos dados originais de citometria, um especialista selecionou uma região de interesse (*ROI – Region of Interest*) para delimitar os dados que foram usados no processo de avaliação. O arquivo original tinha 218.528 células e o conjunto de dados selecionado pelo processo de *gate* apresentava 35.102 células.

A Tabela 1 resume as características dos conjuntos utilizados para a avaliação; isto é, a quantidade de dados de cada conjunto e o número correto de *clusters*. Nos dados simulados o número de grupos é conhecido a priori, enquanto no arquivo de dados de citometria este número foi informado por um especialista. Embora o conjunto de dados de citometria com FL7 x FL8 apresente 35.102 células, o especialista só classificou 30.776 células. Isto aconteceu porque em sua seleção de *gate* nem todas as células foram selecionadas manualmente. Por isso, foi necessário remover 4.326 células deste conjunto.

Tabela 1. Características dos conjuntos de dados utilizados na avaliação.

Conjunto de dados	Quantidade de dados	Num. Clusters
R15	600	15
Compound	399	6
Aggregation	788	7
Citometria (FL7 x FL8)	35.102	3

A Figura 14 mostra os quatro conjuntos de dados e ilustra os diferentes graus de dificuldade impostos a um procedimento automático para estimar o número de *clusters* e também para realizar uma boa inicialização dos centroides de cada *cluster*.

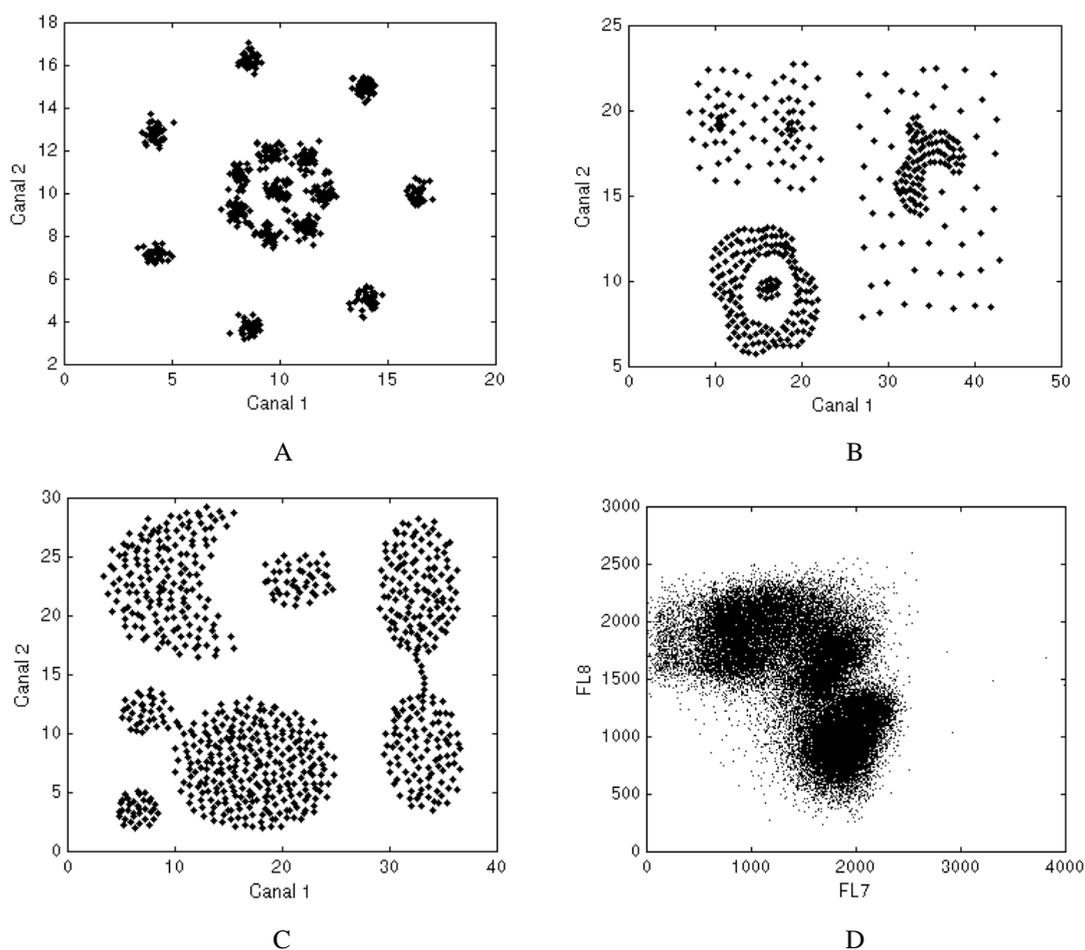


Figura 14. Representação gráfica dos conjuntos de dados. A) R15; B) Compound; C) Aggregation and D) Citometria FL7 x FL8.

A metodologia empregada foi a de submeter cada um dos conjuntos de dados inicialmente à técnica de BIC. Para efeito de comparação, esta metodologia foi comparada com o uso de centroides aleatórios no centro dos dados. Como o uso de

centroides aleatórios já pressupõe que este posicionamento inicial pode não ser o mais adequado, a técnica BIC foi computada cinco vezes para cada quantidade de centroides. A cada rodada seus valores foram registrados guardando-se o menor, o maior e o valor médio da função custo do BIC das cinco tentativas. A cada tentativa, os centroides eram gerados com uma perturbação de 1% em relação ao domínio de cada eixo. Quando foi utilizado o método de modas, a posição inicial dos centroides já fora determinada. Com isso, a técnica BIC somente foi utilizada uma única vez para cada quantidade de centroides e seu valor também foi armazenado. Uma vez que o BIC identifique a quantidade ideal de *clusters* o algoritmo de Mistura de Gaussianas (VEZHNEVETS, *et al.*, 2003) foi acionado para a etapa de agrupamento.

Um programa de computador, chamado de Eflow, foi desenvolvido neste trabalho. Ele possui ferramentas de importação (e exportação) de dados, inicialização de centroides e classificação de dados. Ele foi desenvolvido para suportar dados multidimensionais e seu código está disponível por meio do autor deste trabalho. Este programa foi desenvolvido e testado na plataforma Matlab na versão 7.10, de 64 bits. Optou-se por esta plataforma de desenvolvimento graças a sua facilidade de operação com cálculos matemáticos.

## **V.2 Método de Agrupamento de Dados Quasi-Determinístico Baseado no Coeficiente Individual Kappa (CIK) ou Método Kappa Modificado**

O referido método diz respeito à junção da técnica de inicialização de centroides aleatória no centro médio dos dados, com o emprego do Coeficiente Individual Kappa – CIK (Seção IV.3) para a escolha dos dados de maior relevância, com o método de modelagem e classificação de dados MMG (Seção IV.4). Este método foi aplicado em dados de citometria com o intuito de se investigar uma questão em aberto da área de imunologia, onde ferramentas desenvolvidas foram utilizadas por um especialista em biologia na busca de um novo grupo de células.

A questão biológica investigada está relacionada à maturação de células B do baço. Existe o consenso que estas células B se dividem em três grandes subpopulações: a folicular (FO), a zona marginal (ZM) e a transicional (TR). Outras subdivisões dizem respeito às células transicionais que se subdividem em outros dois subgrupos: uma transicional anterior e outra transicional posterior. Do mesmo modo, as células B FO se

subdividiriam em FO do tipo I e FO do tipo II. Por fim, tem-se ainda a sugestão da existência de uma precursora da zona marginal chamada de pré-zona marginal (PZM). O objetivo da investigação foi encontrar evidências da existência da PZM.

Os dados de citometria de fluxo utilizados nesta etapa do trabalho foram obtidos como descrito a seguir. De um camundongo macho de 8 semanas de idade foram obtidos esplenócitos (células maduras do sistema imunológico) C57BL/6 pelo método de dissociação (GRANATO, 2014). Células em suspensão foram marcadas durante 30 minutos em um citômetro FACS pelos seguintes anticorpos: anti-B220 APC-Cy7 (BioLegend), Anti-IgM Dylight 649 (Jackson ImmunoResearch), anti-CD24 FITC (BioLegend), anti-CD21 PE-Cy7 (BioLegend), anti-CD23 PE (BioLegend), anti-IgD BV605 (BioLegend), e anti-CD93 biotin + streptavidin BV421 (BioLegend). Os dados obtidos do citômetro de fluxo foram tabulados pelo software FlowJo. Um software comercial desenvolvido pela própria empresa FlowJo. Após este, os dados foram exportados para serem trabalhados no Eflow. A metodologia utilizada no Eflow foi a de inserir uma quantidade fixa de centroides inicializados aleatoriamente no centro dos dados e de agrupá-los pelo algoritmo MMG. Devido ao fato do algoritmo MMG necessitar da definição da quantidade de centroides antes de ser iniciado, este foi fixado em seis pelo especialista de citometria. Após a etapa de agrupamento os dados foram novamente exportados em formato de texto (.txt) para serem então trabalhados em um programa gráfico chamado IGOR (Wavemetrics).

A análise de células efetuada pelo citômetro de fluxo apontou a existência de 942.068 células. No entanto, para efeito deste trabalho somente foram utilizadas 300.000 células escolhidas ao acaso. Esta redução de dados é uma prática normal na área de citometria, onde o objetivo é facilitar o trabalho de criação dos *gates* e de posterior interpretação de seus resultados, sem comprometer a análise final. A redução aproximada de 1/3 dos dados foi dada pelo especialista de citometria.

# CAPÍTULO 6

## RESULTADOS

### VI.1 Método de Agrupamento de Dados Determinísticos Baseado em Modas

Os resultados serão apresentados considerando-se quatro aspectos. Inicialmente, serão apresentados os gráficos que demonstram o comportamento dos valores de BIC como função da dimensão de cada modelo. A partir destes gráficos pode-se determinar o melhor número de *clusters* definido pelo método BIC. O segundo resultado indica as posições sugeridas dos centroides pelas duas abordagens. Ora comparando a metodologia clássica do BIC com seus centroides aleatórios no centro dos dados, ora com a técnica de modas juntamente com o BIC. A terceira parte demonstra o tempo gasto por cada método na etapa de agrupamento de dados e a quantidade de iterações realizadas. Por fim, apresentamos o grau de concordância (Coeficiente Kappa) das duas abordagens comparando-os com o padrão ouro. Ambas as abordagens foram submetidas ao algoritmo de MMG em sua etapa de agrupamento de dados. A escolha de se utilizar o algoritmo MMG se baseou no fato de que a expressão da superfície celular apresenta geralmente uma distribuição lognormal (PARKS e HERZENBERG, 1984). Portanto, é razoável assumir que numa escala log, os *clusters* de dados correspondentes a diferentes subpopulações de células devam ter aproximadamente o formato multidimensional de gaussianas.

Mesmo que a abordagem clássica de BIC tenha várias possibilidades de agrupamentos advindos do fato dos centroides serem aleatórios, apenas uma única representação gráfica foi apresentada escolhida ao acaso.

### VI.1.1 Valores de BIC

A Figura 15 apresenta o comportamento dos valores BIC como função da dimensão de cada modelo, isto é, o número de *clusters* dos quatro conjuntos de dados. No lado esquerdo da Figura 15 pode-se observar os resultados do método clássico de BIC. Todos eles têm três linhas representando o valor mais baixo, o valor médio e o valor mais elevado de sua função custo entre todas as cinco tentativas de cada dimensão. Respectivamente as cores que representam esses valores são a vermelha, azul e a preta. Este método clássico define a posição inicial dos centroides no centro dos dados e a cada tentativa era realizada a técnica do *Expectation-Maximization*. Após as cinco tentativas de cada dimensão, esta era incrementada até o seu valor limite de teste. No lado direito da mesma Figura pode-se ver o comportamento do método baseado em modas juntamente com o BIC. Como este método é determinístico tem-se apenas uma única linha (em verde).

Em todas as ilustrações um pequeno círculo azul indica o menor valor de BIC com a sugestão da dimensão do modelo (quantidade de *clusters*).

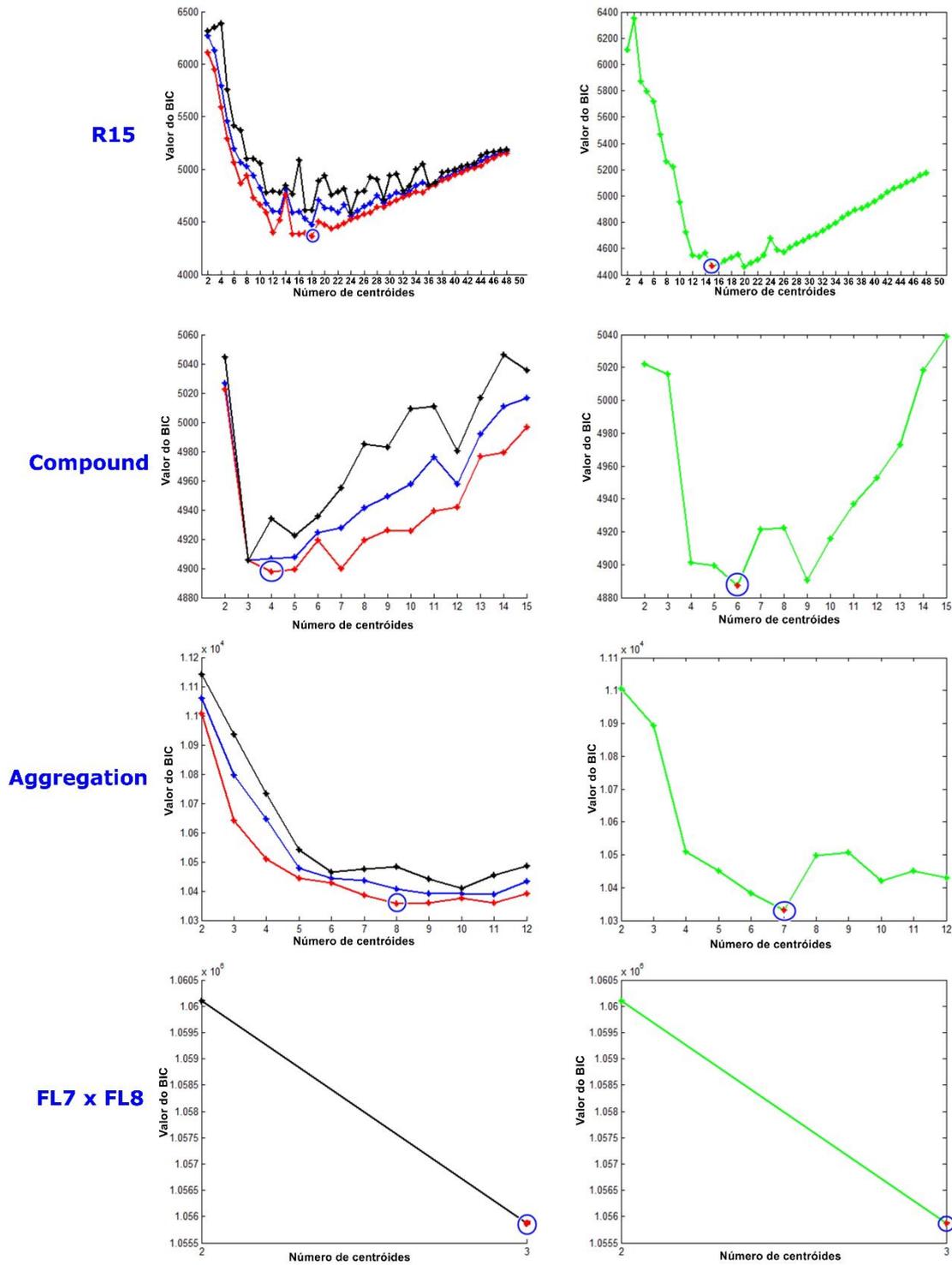


Figura 15. Apresentação gráfica dos valores de BIC (método clássico à esquerda e o método de moda à direita) em função do número de centroides para os quatro conjuntos de dados.

A comparação entre o número real e as quantidades propostas de centroides por cada um dos métodos é apresentada na Tabela 2.

Tabela 2. Número de centroides reais em comparação com as quantidades de centroides propostos pelo método do BIC clássico e pelo método baseado por modas com o BIC.

Conjunto de dados	Número de <i>clusters</i> reais	Número de <i>clusters</i> do método BIC Clássico	Número de <i>clusters</i> do método BIC + Modas
R15	15	18	15
Compound	6	4	6
Aggregation	7	8	7
Citometria (FL7 x FL8)	3	3	3

### VI.1.2 Influência dos Centroides Sugeridos e do Agrupamento Final de Dados

A Figura 16 apresenta as posições iniciais dos centroides sugeridas pelas duas abordagens em questão. Estas posições são ilustradas pelos círculos vermelhos vazados e eles são utilizados pelo algoritmo de MMG para o agrupamento final. A quantidade de círculos representa a quantidade exata de centroides mostrada na Tabela 2.

O resultado final do processo de agrupamento para os três primeiros conjuntos de dados (dados simulados) está representado na Figura 17. Na Figura 18, apresentamos o resultado da amostra de citometria de fluxo quando se utiliza a abordagem clássica do BIC e do método proposto com base em modas assim como também a classificação realizada por um especialista. As cores apresentadas nas Figuras 17 e 18 não representam necessariamente os *clusters* correspondentes para todos os resultados do agrupamento. As cores são escolhidas aleatoriamente pelo programa Eflow.

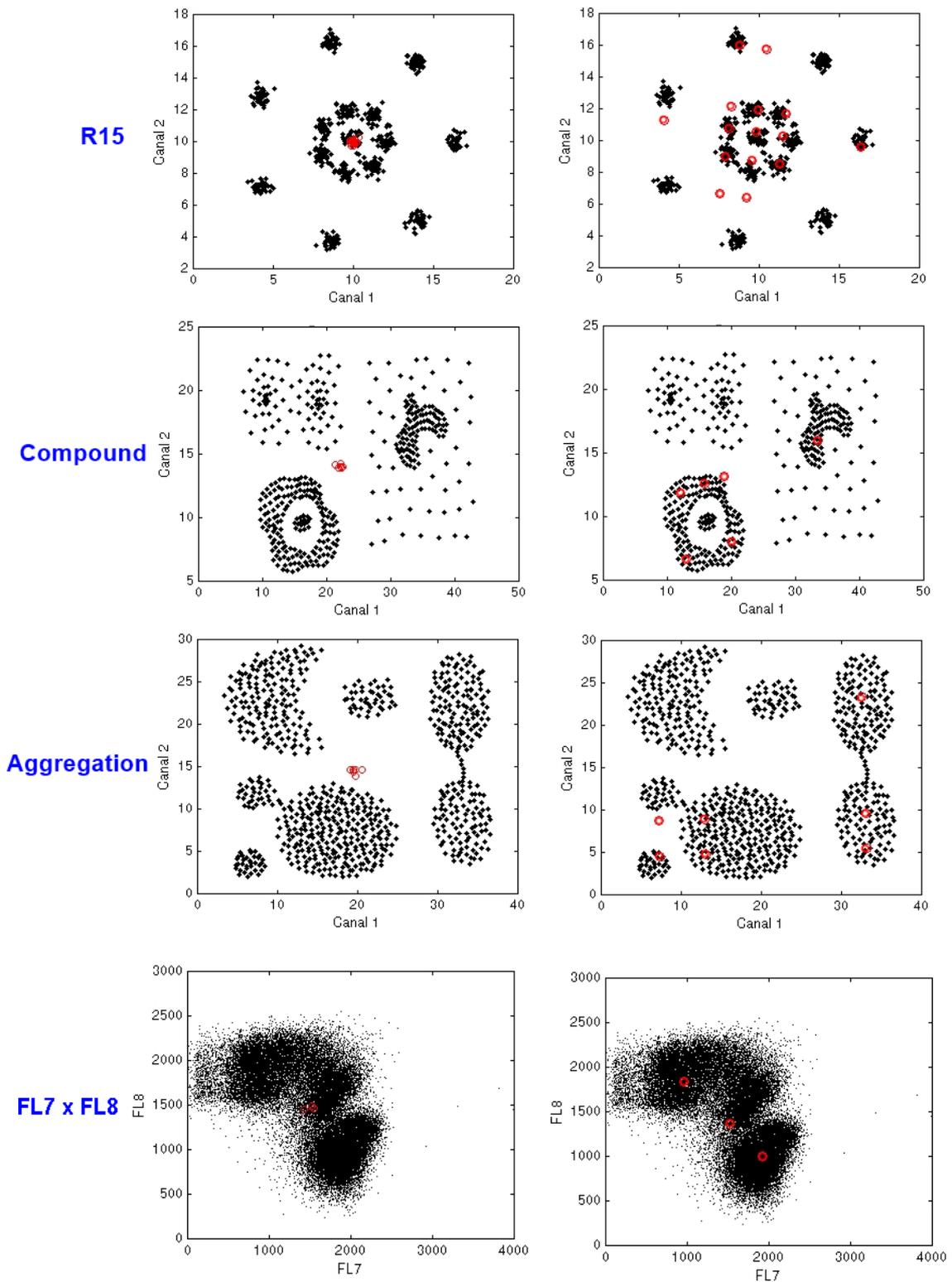


Figura 16. Posições propostas dos centroides pela abordagem clássica do BIC (lado esquerdo) e do algoritmo proposto com base em modas (lado direito). Os centroides estão indicados por círculos vermelhos vazados. Para os gráficos referentes aos arquivos simulados (R15, Compound e Aggregation), os eixos x e y se referem a valores arbitrários para as coordenadas dos dados bidimensionais.

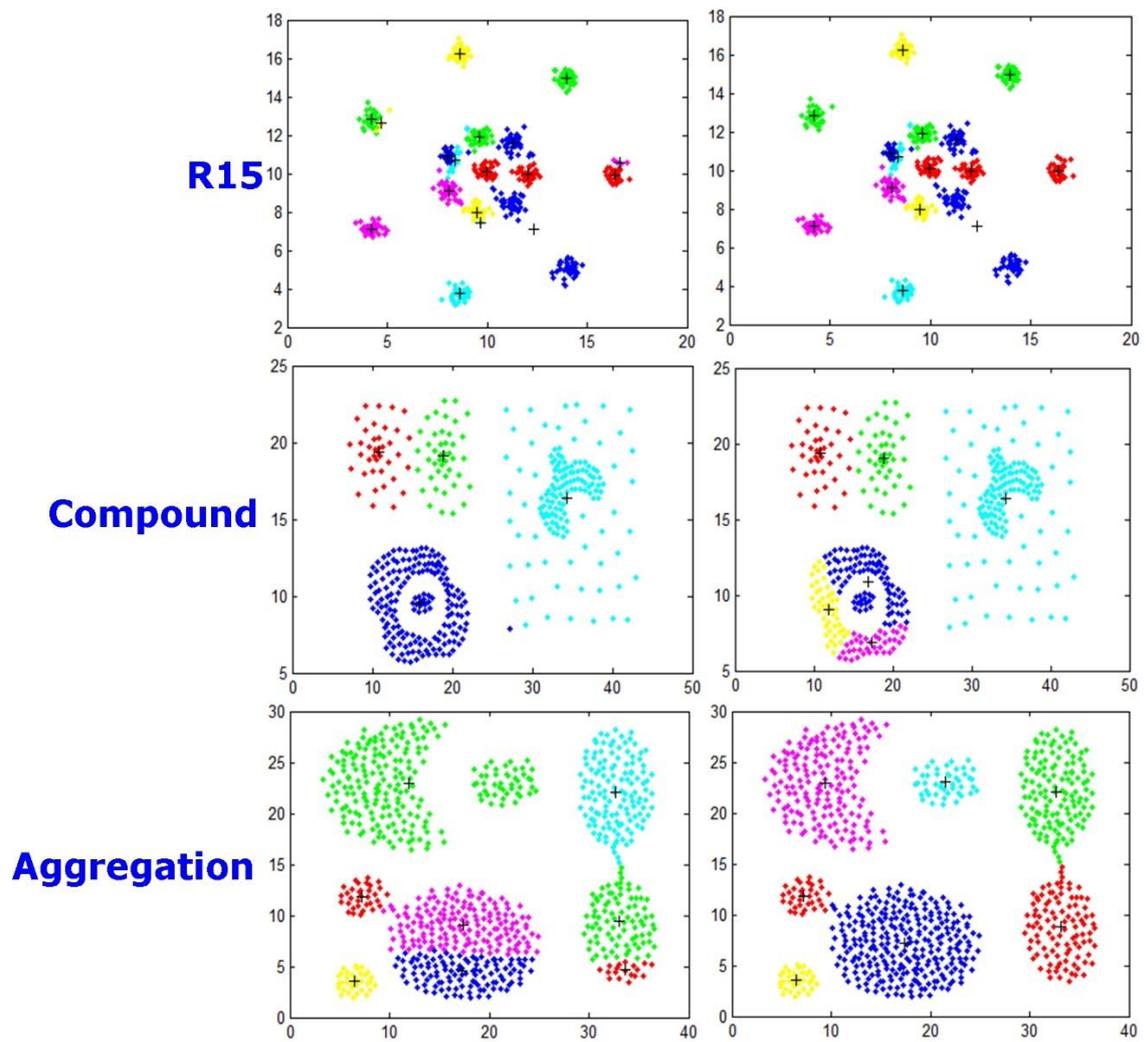


Figura 17. Resultado final do agrupamento para os conjuntos de dados simulados quando se usa a abordagem clássica do BIC (lado esquerdo) e do algoritmo baseado em modas (lado direito). Os eixos x e y se referem a valores arbitrários para as coordenadas dos dados bidimensionais

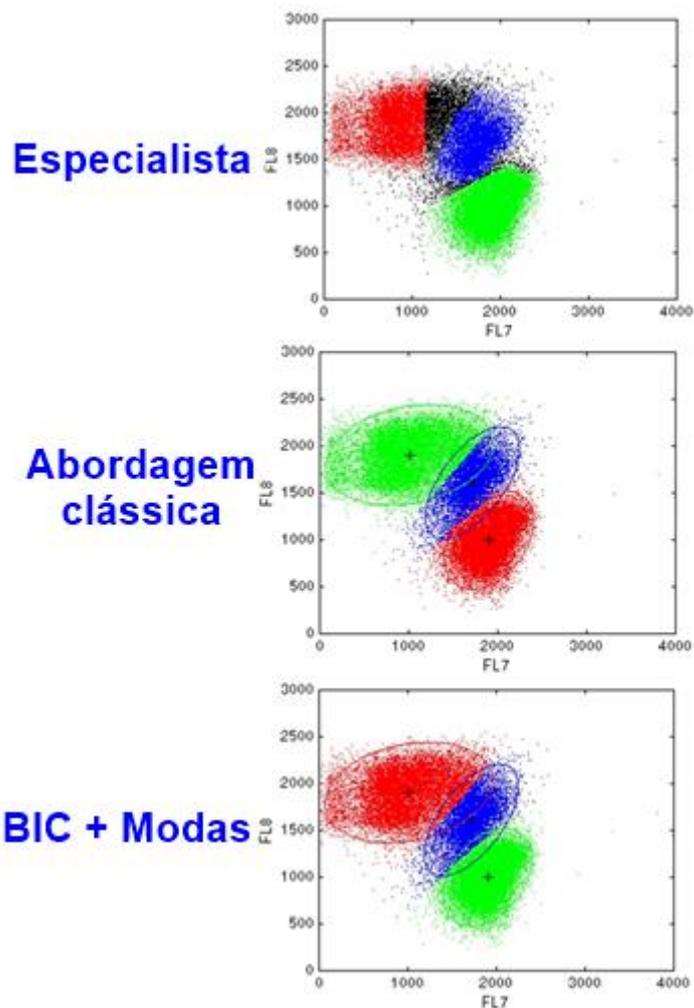


Figura 18. Resultado final do agrupamento para o conjunto de dados de citometria de fluxo realizada pelo especialista (acima), quando se usa a abordagem clássica (meio), e quando se usa o algoritmo baseado em modas (abaixo).

### VI.1.3 Tempo Gasto e Iterações

Também foi avaliado o tempo gasto de cada processo de agrupamento e a quantidade de iterações necessárias para cada conjunto de dados considerando-se as duas abordagens. Ambas as medições foram resumidas e estão apresentadas na Tabela 3. Todas essas medições foram realizadas em um computador com Matlab 7.10, processador Intel I3 de 2,13 GHz e 8,00 GB de memória RAM.

Tabela 3. Número de iterações e tempo gasto para a etapa de agrupamento para o método de BIC clássico e para o método baseado em modas.

Conjunto de dados	BIC Clássico		BIC + Modas	
	Quantidade de iterações	Tempo para agrupamento (s)	Quantidade de iterações	Tempo para agrupamento (s)
R15	44	0,9464	44	0,8906
Compound	65	0,3983	119	0,6632
Aggregation	283	1,4769	127	0,8007
Citometria (FL7 x FL8)	91	3,8051	91	3,5651

#### VI.1.4 Concordância com o padrão ouro

Os dados da Tabela 4 têm como objetivo medir o grau de concordância entre os agrupamentos em relação ao padrão-ouro. O coeficiente Kappa (K) tem um intervalo de confiança de 95%.

Tabela 4. Comparação dos coeficientes de Kappa (K) entre o método clássico do BIC e o método baseado em modas. Ambos utilizando a técnica de MMG para seu agrupamento.

Conjunto de dados	BIC Clássico		BIC + Modas	
	K	Variância	K	Variância
R15	87,64%	$< 10^{-5}$	89,20%	$< 10^{-5}$
Compound	77,64%	0,0007	50,33%	0,0004
Aggregation	52,14%	0,0004	99,81%	0,0005
Citometria (FL7 x FL8)	93,33%	$< 10^{-5}$	93,33%	$< 10^{-5}$

## **VI.2 Método de Kappa Modificado para Definição da Pré-Zona Marginal**

Antes de se iniciar esta seção de resultados sobre a investigação da pré-zona marginal utilizando-se o método Kappa modificado, é importante mencionar que o método de agrupamento de dados determinísticos baseado em modas foi testado na amostra biológica, mas não foi possível para o algoritmo encontrar mais do que um centroides. Em todas as seis projeções das componentes principais apenas viu-se uma distribuição unimodal.

### **VI.2.1 Classificações Iniciais**

Os seis centroides inicializados aleatoriamente em torno do centro dos dados foram submetidos ao algoritmo de mistura de gaussianas. Como esta geração de centroides é aleatória e não há qualquer garantia que duas, ou mais, inicializações das posições desses centroides pudessem resultar no mesmo agrupamento final, esta rotina foi realizada dez vezes. Um especialista em citometria analisou, então, cada uma dessas dez classificações, e, segundo ele, foi possível identificar dois padrões de classificações. A identificação foi apenas visual e baseada em sua experiência. Deste modo, ele dividiu as dez classificações em dois grupos, sendo um com quatro das dez classificações e o outro com seis classificações. Estes dois grupos foram denominados de agrupamento A e A', respectivamente. Como as classificações de cada Agrupamento (A e A') eram semelhantes entre si, uma classificação de cada agrupamento foi selecionada ao acaso para efeito da análise que se segue. Considerando-se, então, uma das classificações de cada agrupamento A e A', pode-se observar na Tabela 5 a quantidade de células de cada um dos *clusters*, juntamente com o seu percentual em relação ao total de 300.000 células utilizadas nesta etapa do trabalho.

Tabela 5. Apresentação do total de células em cada *cluster* para os agrupamentos A e A' de uma classificação de cada tipo escolhida ao acaso, bem como sua distribuição percentual em relação ao total de 300 mil células.

<i>Cluster</i>	<b>Agrupamento do tipo A</b>		<b>Agrupamento do tipo A'</b>	
	<b>Total células</b>	<b>% células</b>	<b>Total células</b>	<b>% células</b>
1	42.830	14,28%	38.304	12,76%
2	15.367	5,12%	32.476	10,83%
3	67.310	22,44%	64.762	21,59%
4	79.847	26,61%	72.048	24,02%
5	54.519	18,17%	56.786	18,93%
6	40.127	13,38%	35.624	11,87%

A Figura 19 apresenta as principais vistas 2-D para as 300.000 células. Em cada linha da Figura estão os seis *clusters* (de C1 até C6) do agrupamento A da classificação escolhida ao acaso. Em cada coluna encontram-se as combinações das principais projeções canônicas associadas ao experimento realizado. As principais subpopulações da célula B foram relacionadas com os *clusters* conforme a Figura.

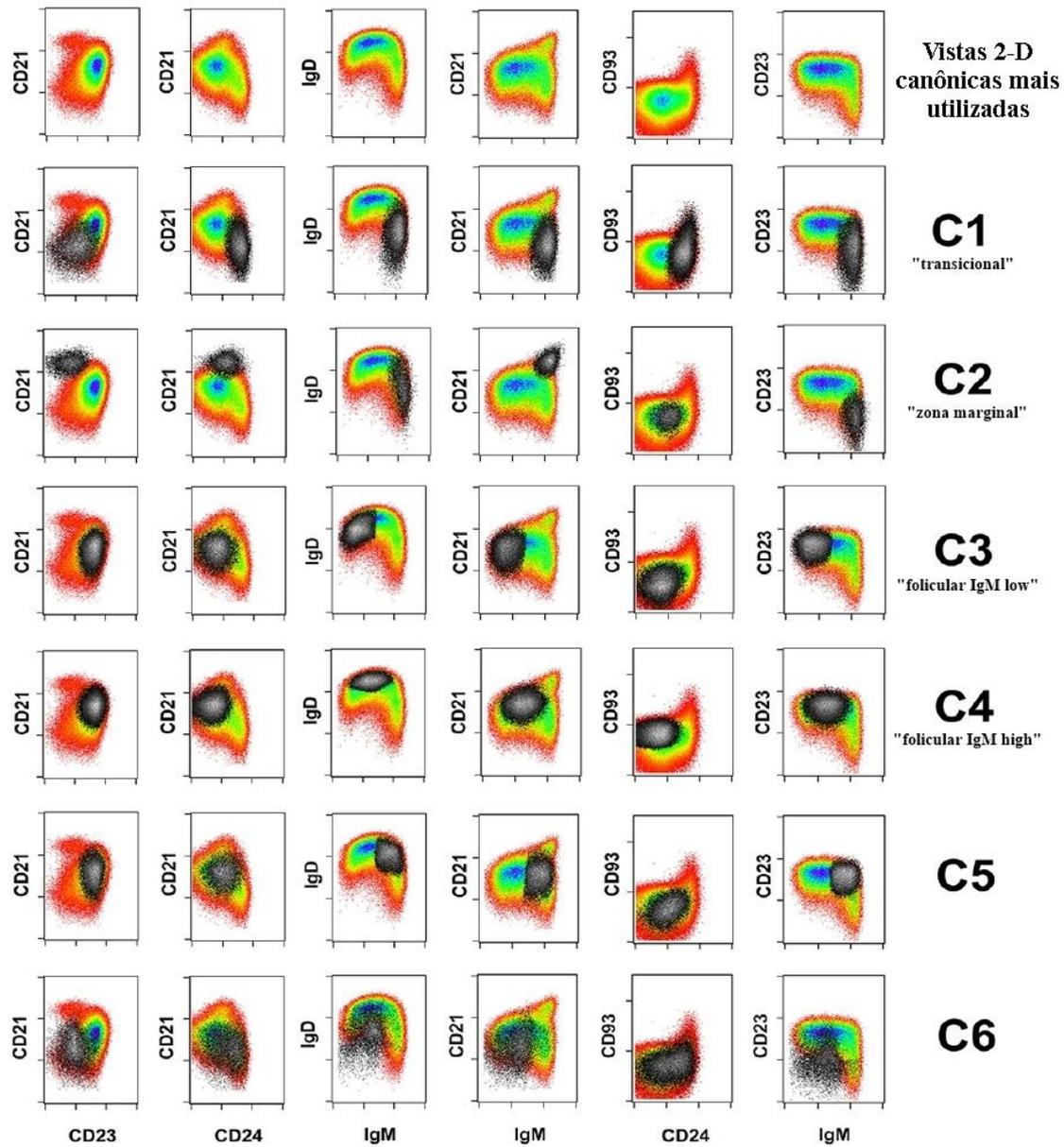


Figura 19. Vistas canônicas 2-D das 300 mil células. Cada linha de imagens, após a primeira linha, representa os agrupamentos realizados pelo EFlow, indo de C1 a C6, referentes ao agrupamento do tipo A. Os *clusters* estão representados por pontos pretos.

É possível notar as seguintes subpopulações: C1 => transicional; C2 => zona marginal; C3 => folicular  $IgM^{low}$ ; C4 => folicular  $IgM^{high}$ ; C5 para  $IgD^{high}$  e  $IgM^{bright}$  com um fenótipo menos evidente; e C6, contendo a maior parte de células  $IgD^{low}$  e  $IgM^{low}$ , com um fenótipo não familiar.

A Figura 20 mostra os seis *clusters* (C1' a C6') definidos por uma das classificações escolhida ao acaso do agrupamento A'.

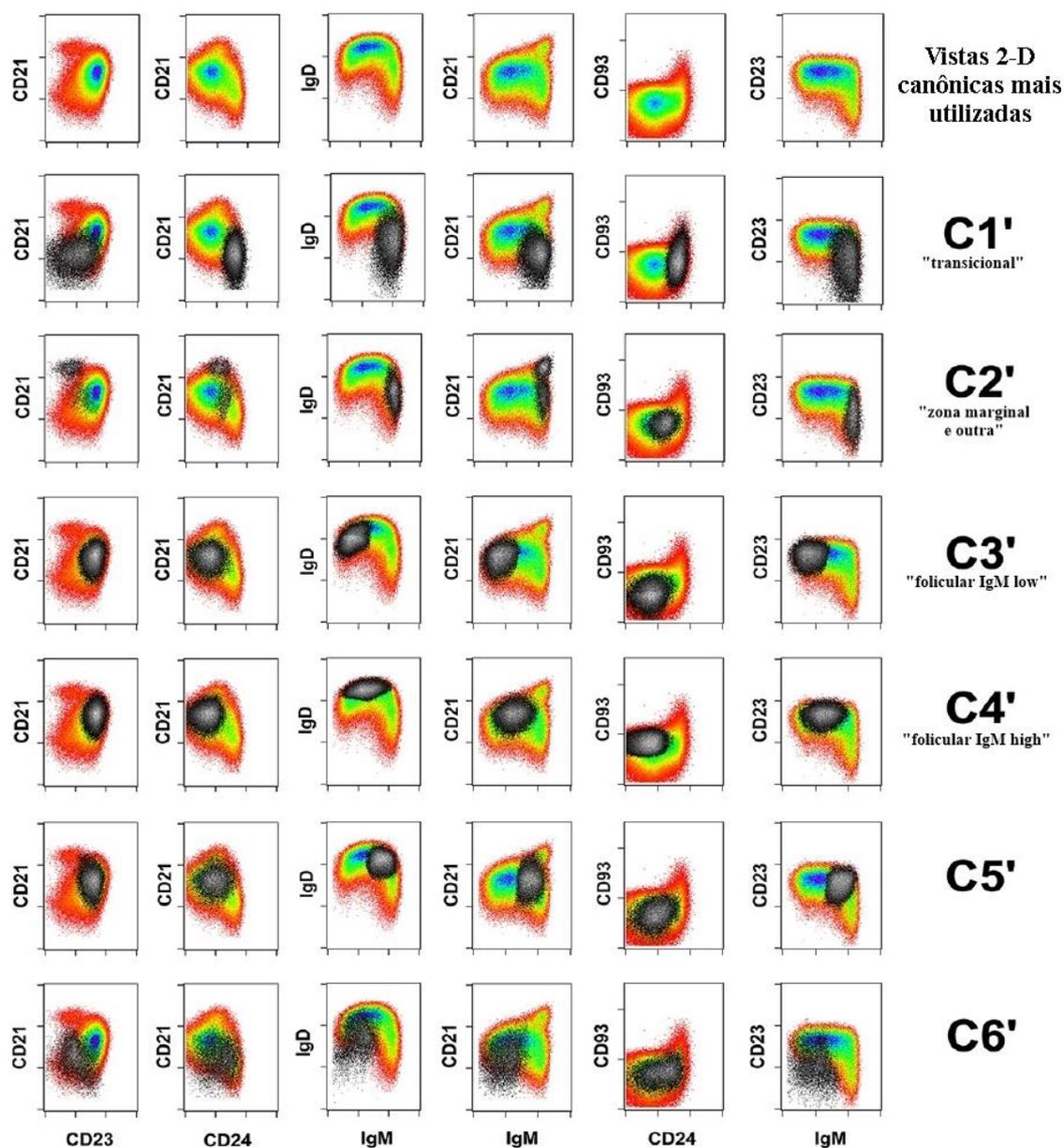


Figura 20. Vistas canônicas 2-D das 300 mil células. Cada linha de imagens, após a primeira linha, representa os agrupamentos realizados pelo EFlow, indo de C1' a C6', referentes ao agrupamento do tipo A'. Os *cluster* estão representados por pontos pretos.

Após a identificação dos seis *clusters* associados aos dois agrupamentos (A e A'), o método proposto de Kappa modificado foi aplicado às classificações de cada tipo de agrupamento. Ou seja, para o agrupamento A o Kappa modificado foi submetido para as suas quatro classificações; enquanto que para o agrupamento do tipo A', o método foi aplicado as suas seis classificações. Foi estabelecido um CIK igual 1 (o mais rigoroso possível) para ambos os casos. O resultado de cada análise está resumido na Tabela 6.

Tabela 6. Apresenta o quantitativo original de células, o quantitativo de células após o uso do CIK e o percentual de células reduzidas, para ambos os agrupamentos A e A'.

<b>Agrupamento A</b>			
<i>Cluster</i>	<b>Células originais</b>	<b>Células com CIK=1</b>	<b>Redução (%)</b>
C1	42.830	42.812	0,04%
C2	15.367	15.362	0,03%
C3	67.310	67.271	0,06%
C4	79.847	79.768	0,10%
C5	54.519	54.405	0,21%
C6	40.127	40.111	0,04%
<b>Total</b>	<b>300.000</b>	<b>299.729</b>	<b>0,09%</b>
<b>Agrupamento A'</b>			
C1'	38.304	36.619	4,40%
C2'	32.476	28.884	11,06%
C3'	64.762	39.385	39,19%
C4'	72.048	59.477	17,45%
C5'	56.786	45.074	20,62%
C6'	35.624	25.601	28,14%
<b>Total</b>	<b>300.000</b>	<b>235.040</b>	<b>21,65%</b>

Buscando exemplificar visualmente o impacto da redução de células, a Figura 21 apresenta o *cluster* C2' antes e depois do método de Kappa modificado. Este *cluster* foi o escolhido porque ele será objeto de estudo na subseção seguinte.

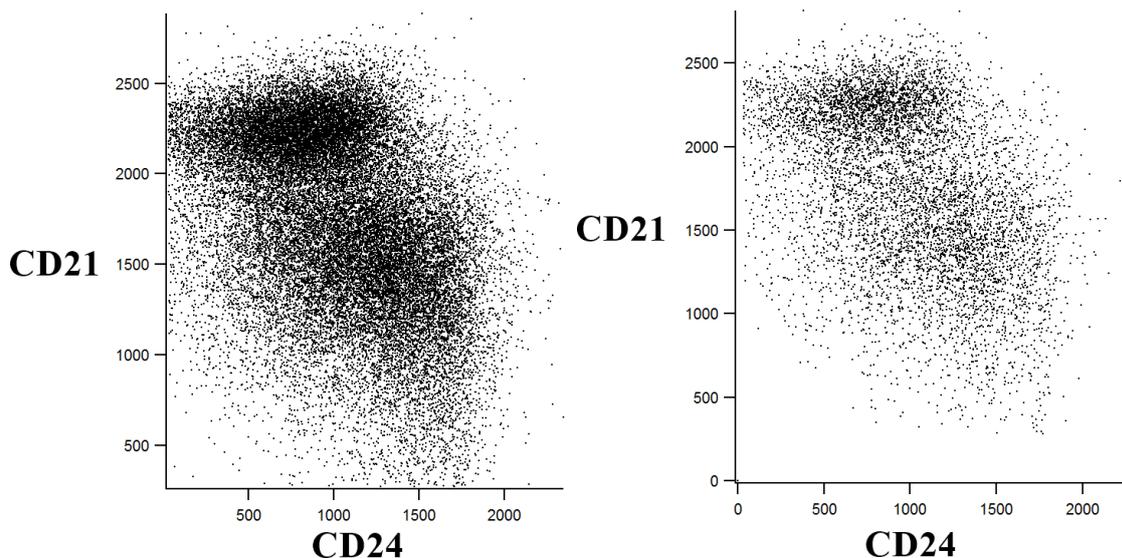


Figura 21. A Figura do lado esquerdo apresenta as células do C2' em sua integridade, com 32.476 células. Na Figura do lado direito a mesma vista, CD24 x CD21, apresenta o mesmo *cluster* mas com uma redução 11,06%.

### VI.2.2 Pré-Zona Marginal

Muitos dos *clusters* do agrupamento A' (C1', C3', C4', C5' e C6') são semelhantes aos *clusters* do agrupamento A, excetuando-se o *cluster* C2'. Este *cluster* C2' contém o *cluster* C2 e mais algumas outras células que parecem formar um novo subgrupo. Por isso este *cluster* C2' foi submetido novamente ao EFlow para ser subdividido em outros dois novos *clusters*. Este novo agrupamento de dados está apresentado na Figura 22.

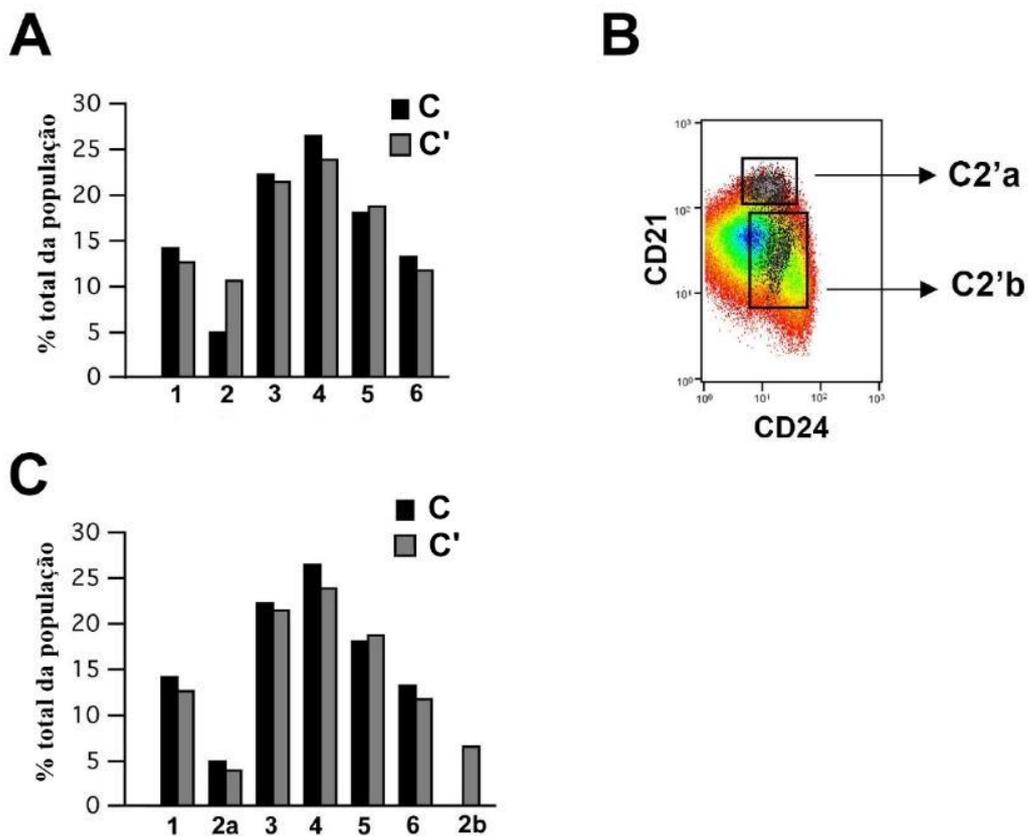


Figura 22. A Figura A apresenta um histograma do total de células (em valores percentuais) para cada um dos 6 *clusters* dos grupos C e C'. A Figura B informa a localização dos *clusters* C2'a e C2'b nas vistas CD24 x CD21 com pontos na cor preta. A Figura C apresenta um histograma do total de células (em valores percentuais) dos grupos C e C' com destaque para o novo subgrupo 2b do grupo C'.

A Figura 22A apresenta um histograma com o percentual de células em cada *cluster* para os agrupamentos A e A'. A Figura 22B apresenta a vista CD24 x CD21 com a identificação do *cluster* C2' subdividido em: C2'a e C2'b. O *cluster* C2'a corresponde ao *cluster* C2 do agrupamento A. Isto pode ser confirmado quando são observados os percentuais de células de ambos os *clusters* apresentados no histograma da Figura 22C. Neste histograma é possível notar que os *clusters* C2 e C2'a se aproximam na quantidade de células, deixando as células C2'b agora separadas.

Para comparar a metodologia clássica de exploração e análise de dados de um especialista contra o método de Kappa modificado nesta subpopulação de C2', submetemos este *cluster* a nove novas classificações. Dois centroides aleatórios foram posicionados no centro dos dados e em seguida o algoritmo de MMG foi aplicado. O EFlow exporta como resultado um arquivo representando uma matriz de dados N x M onde N é o número de células analisadas e M indica o número de *clusters* a que cada célula pode pertencer (neste caso apenas os valores 1 ou 2). Cada um dos nove resultados

gerou um arquivo N x M. Juntamente com o arquivo original, os nove novos resultados foram submetidos a uma planilha de dados que calculou o coeficiente Kappa, sendo o coeficiente desejado igual a 1. Ao aplicar um CIK igual a um, todas as 32.476 células foram selecionadas. A Figura 23A apresenta o *cluster* C2' original, e na Figura 23B os dois *clusters* formados após a etapa de agrupamento do algoritmo de MMG. Como o coeficiente Kappa das dez classificações foi igual a 1, isto significa dizer que CIK de todas as células também foi igual a 1. Desta forma, não houve redução de células na Figura 23B com o método do Kappa modificado.

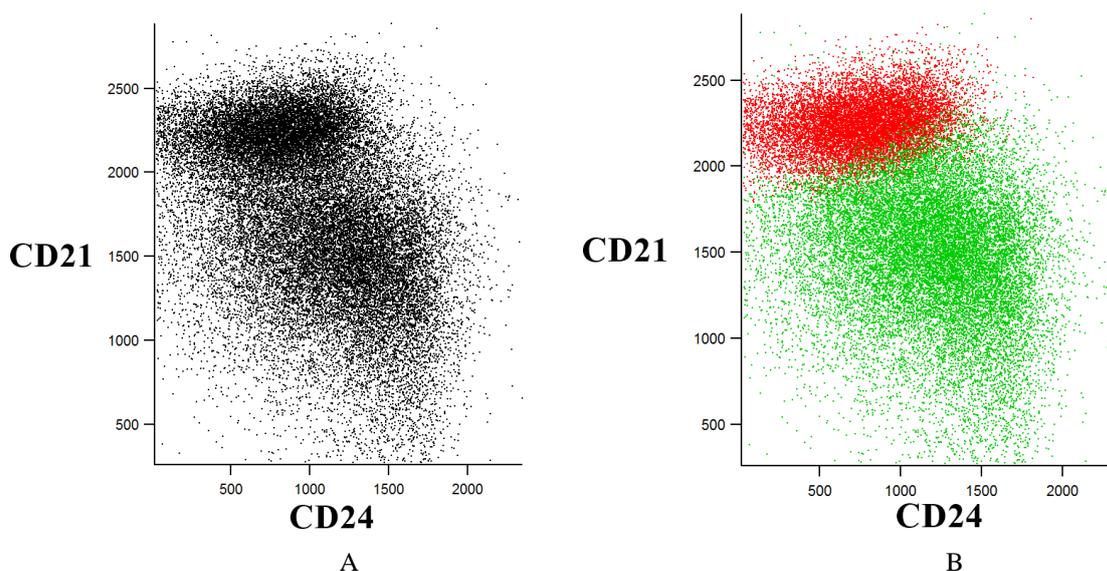


Figura 23. A Figura A apresenta o *cluster* C2' original, e na Figura B os dois *clusters* formados após a etapa de agrupamento do algoritmo de MMG nas vistas CD24 x CD21.

### VI.2.3 Eixo Separador

A partir dos *clusters* obtidos foi realizado o procedimento padrão de agrupamento usado na área de análise de dados de citometria que se baseia na inspeção visual de várias projeções canônicas, buscando separar as populações clássicas principais: TR, FO e ZM. Esses três grandes grupos correspondem a cerca de 66% do total de células e são relativos aos *clusters* C1', (C3' + C4') e C2'a, respectivamente. A Figura 24 mostra as vistas CD24 x CD21 e CD23 x CD21 apenas com as nuvens de pontos (e gráficos de contorno) desses três grandes grupos.

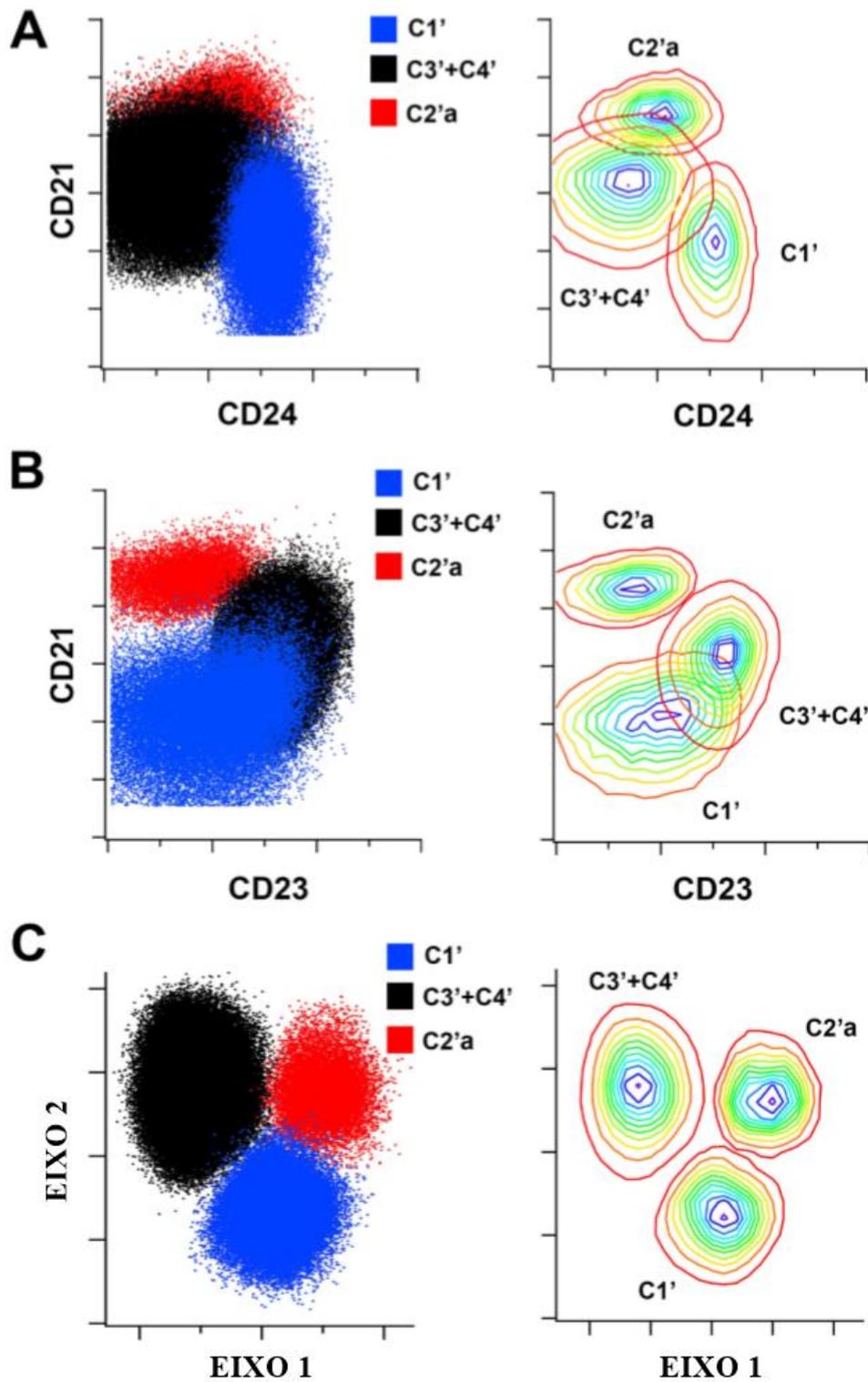


Figura 24. As Figuras do lado esquerdo representam as projeções dos *clusters* C1', C3'+C4' e C2'a, nas vistas a) CD24 x CD21, b) CD23 x CD21 e c) Eixo 1 x Eixo 2 determinadas empiricamente. Os *clusters* C1', C3'+C4' e C2'a representam as células transicionais (azul), folicular (preto) e zona marginal (vermelho), respectivamente. As Figuras do lado direito denotam os mesmos *clusters* e as projeções do lado esquerdo, mas agora representados em curvas de nível.

Graças às ferramentas gráficas do software IGOR os mesmos três *clusters* foram rotacionados entre si em busca de uma projeção (Eixo 1 x Eixo 2) que fugisse das projeções canônicas tradicionais. O resultado pode ser observado na Figura 24C e deve ser mencionado que o plano de projeção não canônico, formado pelos Eixos 1 e 2 mostrados na Figura, foi encontrado de forma empírica, modificando-se manualmente os eixos canônicos no programa Igor até uma posição de interesse. As combinações lineares das projeções do Eixo 1 e 2 são:

- Eixo 1 = 1.0 [IgM] – 0.6 [IgD] – 0.5 [CD23];
- Eixo 2 = 1.0 [CD21] – 1.0 [CD24].

As escalas dos eixos estão em unidades arbitrárias.

A partir dos Eixos 1 e 2 foram também inseridos os *clusters* C2'b e C5'. A Figura 25 apresenta essas projeções.

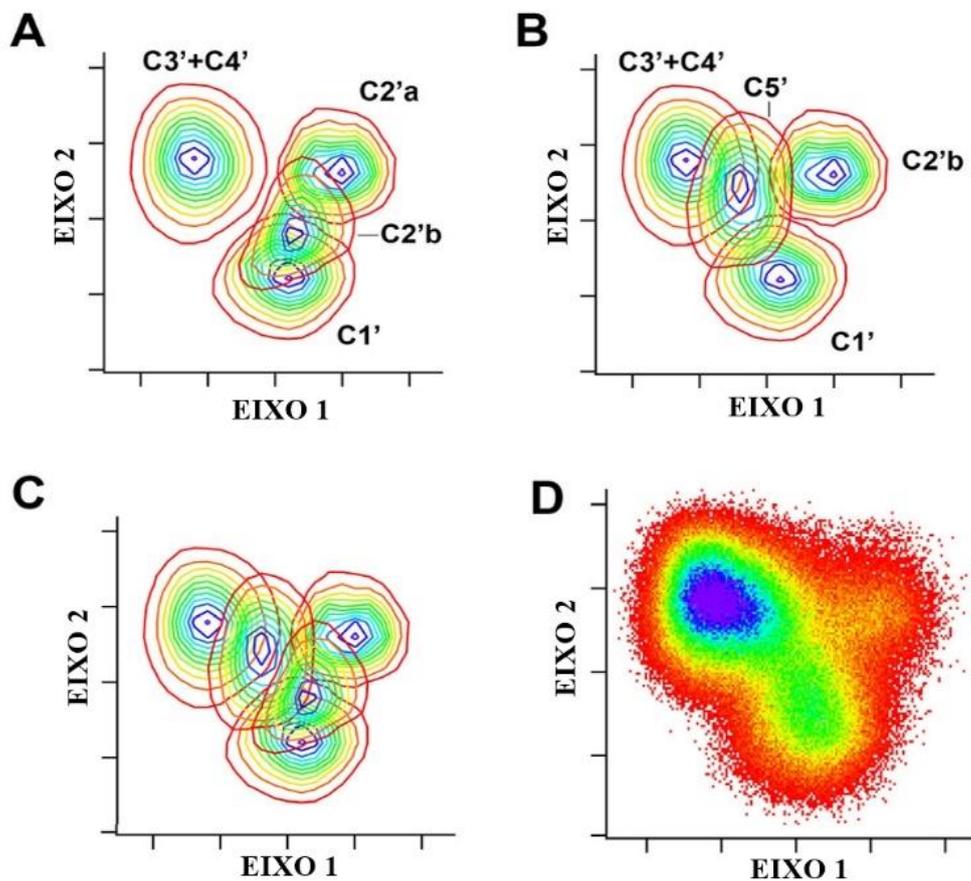


Figura 25. Apresentação dos *clusters* C1', C3'+C4' e C2'a com o(s) *cluster*(s): (A) C2'b; (B) C5'; (C) C2'b e C5'. As Figuras A, B e C são representadas por curvas de nível. A Figura D por densidade de pontos.

# CAPÍTULO 7

## DISCUSSÃO

Este capítulo foi subdividido em duas secções para melhor entendimento de cada assunto. A primeira seção refere-se ao método de inicialização baseado em modas e a segunda seção refere-se ao método de Kappa modificado.

### VII.1 Discussão sobre o Método de Centroides Baseados em Modas

As classificações para os conjuntos de dados simulados apresentaram um melhor resultado quando se usou o algoritmo de inicialização baseado em modas. Os resultados da Tabela 2 apontam que em todos os conjuntos de dados, a quantidade correta de centroides foi encontrada pelo algoritmo baseado em modas, contra todas as abordagens do BIC clássico. É importante reforçar que, em todas as suas tentativas, a inicialização aleatória no centro dos dados da abordagem clássica do BIC sugeriu uma quantidade de centroides diferente do correto. A única exceção ocorreu com o conjunto de dados de citometria. Como a Tabela 2 apresenta apenas uma única tentativa, este tipo de resultado pode mudar a quantidade de acertos possíveis.

O segundo ponto a ser ressaltado é sobre a análise gráfica. No conjunto de dados R15 a classificação foi, visualmente, muito boa em ambas as abordagens. No conjunto de dados Compound, a abordagem clássica BIC parece muito melhor que abordagem por modas. No entanto, nesse conjunto de dados, a abordagem clássica BIC não pode determinar o número certo de centroides. Este conjunto de dados pode ser considerado difícil de ser agrupado porque suas formas são difíceis de serem identificadas por métodos não-supervisionados. Sobre o conjunto Aggregation, o método de agrupamento por modas foi quase perfeito. Na abordagem do BIC clássico apenas três grupos foram

corretamente classificados. Já na análise visual dos dados de citometria, a *clusterização* foi assemelhada para ambas as abordagens.

No primeiro gráfico (superior) da Figura 18, três regiões foram selecionadas (vermelho, azul e verde). Os pontos na cor preta representam células que não foram selecionadas pelo especialista. Assim, eles não foram considerados no teste de comparação. No que diz respeito ao conjunto de dados de citometria mostrados nas projeções FL7 x FL8, não há uma diferença visível entre os resultados dos dois métodos de inicialização. O mesmo foi observado para os valores de Kappa, como veremos a seguir.

Sobre o número de iterações e o tempo computacional ambos são correlacionados. O conjunto de dados R15 e a amostra de citometria tiveram o mesmo número de iterações entre si nos dois métodos e um desempenho no aspecto tempo ligeiramente melhor no método baseado em modas. Para o conjunto Compound a abordagem clássica de BIC apresentou um número menor de iterações que o método baseado em modas, convergindo naturalmente mais rápido que seu concorrente. No entanto, no método BIC clássico, apenas quatro centroides foram sugeridos quando deveriam ter sido sugeridos seis. Com menos *clusters* é natural que o tempo de processamento para um agrupamento seja menor, quando comparado com a mesma amostra com um número maior de centroides a classificar. O objetivo não é apenas obter menos iterações e ser mais rápido, mas sim se há alta concordância com o padrão ouro com melhor redução de tempo e custo computacional. Observando-se a Figura 17, é fácil de se entender que a concordância não foi obtida porque o conjunto de dados original possui seis grupos, ao invés de quatro. No método baseado em modas, a concordância não foi boa, 50,33% (Tabela 4), mesmo tendo proposto o número certo de centroides. Isto significa que ainda é importante estudar métodos de agrupamento de dados com boas técnicas para a inicialização centroides.

Para a abordagem baseada em modas, o coeficiente Kappa do conjunto Aggregation teve o melhor grau de concordância, 99,81%; enquanto o conjunto de dados Compound apresentou o resultado pior, 50,33%. Isso significa que, apesar do método de modas melhorar a qualidade de agrupamento, esta técnica não é a melhor em qualquer situação. Dependendo das características de agrupamento dos dados, a técnica tem que ser adaptada ou alterada.

Concluindo, foi apresentado um novo algoritmo de inicialização de centroides baseado em modas das projeções dos dados nas componentes principais, o qual produziu bons resultados quando utilizado em conjunto com o algoritmo MMG. É importante notar que nenhuma informação a priori de dados foi necessária, resultando num tipo de técnica não supervisionada multidimensional. Sua vantagem está em produzir resultados determinísticos, permitindo ao usuário fazer análises repetitivas e reprodutíveis, sem necessidade de novas classificações.

## VII.2 Discussão sobre o Método de Kappa Modificado

A Figura 21 (imediatamente antes da subsecção VI.2.2 Pré Zona Marginal) apresenta o *cluster* C2' antes e depois da aplicação do Kappa modificado. Pode-se perceber que a Figura do lado direito se apresenta bem mais clara que a mesma Figura do lado esquerdo. Sugerindo, em um primeiro momento, que a redução de células tenha sido expressiva. No entanto, esta redução foi de apenas 11,06%, conforme informado na Tabela 6. Esta situação representa que análises puramente visuais podem levar o usuário a uma impressão errônea sobre os dados apresentados. Outro exemplo bastante comum, e que podem levar citometristas a resultados diferentes, é a utilização de *gates*. Dependendo da experiência prévia do citometrista e de outros aspectos subjetivos, é possível se obter diferentes resultados de análise, mesmo quando a amostra de dados é a mesma. Situações como estas acontecem exatamente pela interpretação subjetiva de cada especialista, e sua habilidade visual é um dos parâmetros para essas interpretações; mas que eventualmente podem estar equivocadas. O estabelecimento de uma padronização no método usado para realizar o agrupamento dos dados diminuiria diferenças entre especialista e também de interpretações visuais.

Acreditamos que existam dois grandes motivos para que as dez classificações geradas a partir de C2', após o Kappa modificado com  $CIK=1$ , tenham sido exatamente iguais. O primeiro deles seria a correta escolha da quantidade de centroides. Ou seja, de que realmente existem dois, e somente dois, *clusters* muito bem definidos. O segundo motivo refere-se à possível boa localização inicial dos centroides. Como o passo dado pelo centroide do algoritmo de clusterização é do tipo Gradiente Descendente, as perturbações máximas de 1% no domínio dos dados não são suficientes para que as posições dos centroides “fujam” de mínimos locais, ou global. Caso a função custo pudesse rejeitar boas opções de localização dos centroides (de menor custo), para testar

outras posições (de maior custo), seria possível testar a hipótese de que as posições iniciais desses centroides no centro dos dados não fossem mínimos locais. Uma outra forma de avaliar se o centro dos dados é um mínimo global, seria realizando uma comparação com um padrão ouro. Para isso, um teste *in vitro* poderia confirmar tal suposição com o uso de um *sorter*. Como os dez agrupamentos foram 100% iguais acreditamos se tratar de um mínimo global.

Outro fato muito importante diz respeito ao significado biológico do C2'b. Sua representação biológica pode quebrar um paradigma sobre o real comportamento de amadurecimento celular das células B antes de pertencerem à zona marginal. As células C2'b tem seu fenótipo similar às células B da ZM e da transicional, como pode ser visto na Figura 26.

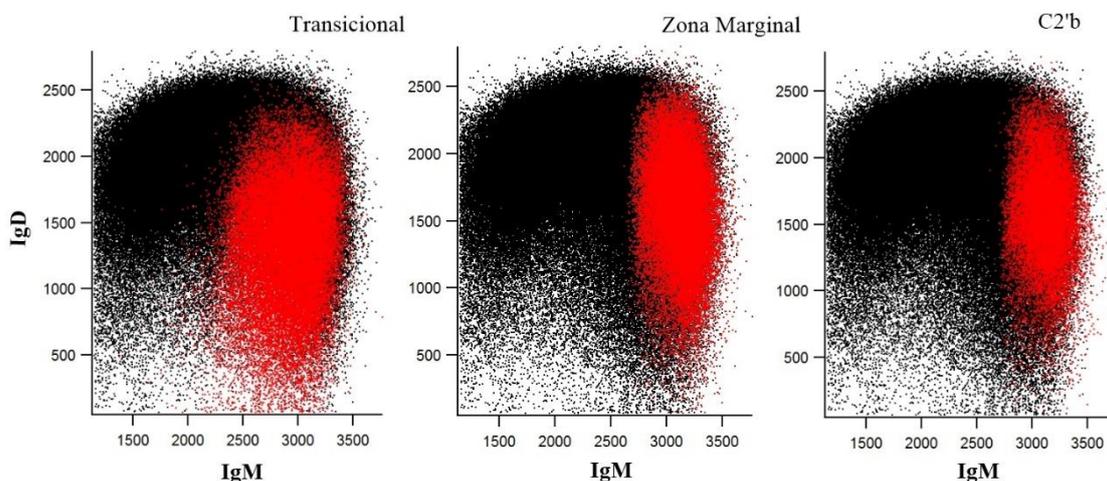


Figura 26. Nas vistas IgM x IgD são apresentadas as 300 mil células com destaque na cor vermelha para os grupos: Transicional; Zona Marginal; e *cluster* C2'b, nesta ordem.

No entanto, o *cluster* C2'b possui uma menor expressão de CD24 do que as células B Transicionais, diferindo fenotipicamente dessas. Isto pode ser observado na vista CD24 x CD21 da Figura 27. Mesmo que a nuvem de pontos vermelha se apresente extensa, C2'b apenas representa cerca de 6,7% do total de pontos. As células Transicionais (em verde e à direita) representam cerca de 12,2%. Ou seja, as células C2'b não são células B Transicionais em sua essência.

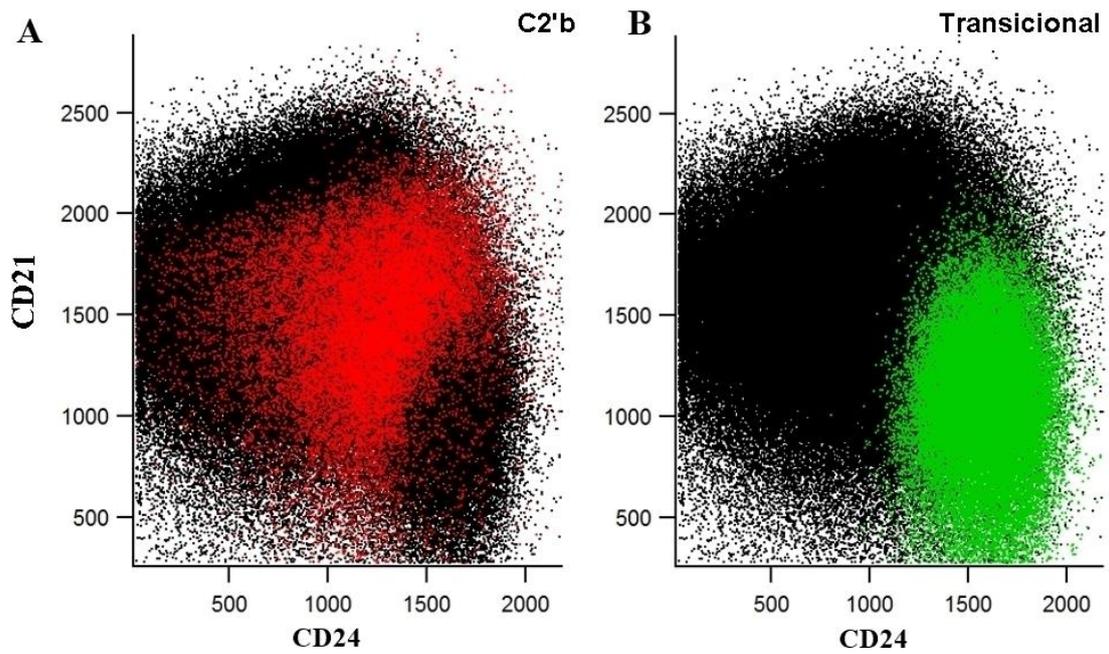


Figura 27. Nas vistas CD24 x CD21 são apresentadas as 300 mil células com destaque para: (A) *cluster* C2'b em vermelho; e (B) células Transicionais em verde.

Continuando o raciocínio, embora o *cluster* C2'b esteja relacionado fenotipicamente com a Zona Marginal, conforme pode ser visto na Figura 28, as células C2'b não são CD21<sup>high</sup> quando são observadas pela vista CD23 x CD21. A Figura 28 apresenta a falta de CD21 nas células C2'b.

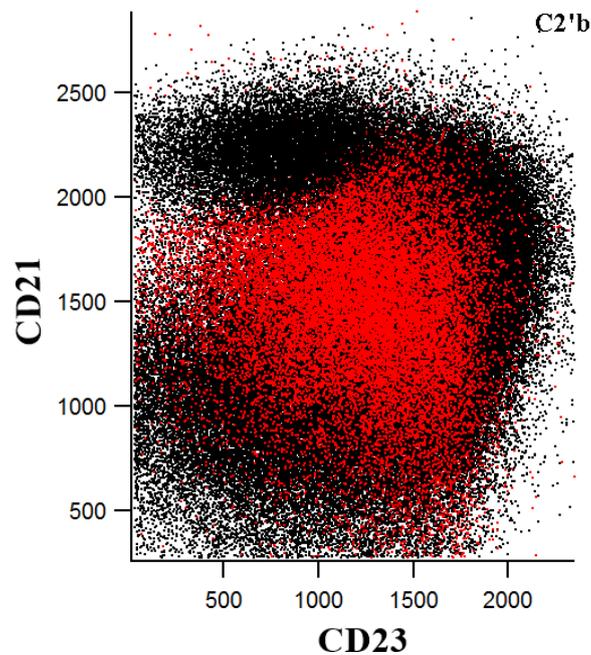


Figura 28. Os pontos vermelhos representam as células classificadas no *cluster* C2'b na vista CD23 x CD21 em relação ao total de células.

Por isso, as células pertencentes ao *cluster* C2'b não poderiam ser classificadas como PZM, de acordo com um dos métodos clássicos da literatura (SRIVASTAVA, B. *et al.*,2005) para a classificação de uma PZM, onde a ZM e a PZM são CD21<sup>high</sup>. Na Figura 29 estão apresentadas as posições clássicas de células para serem classificadas como uma ZM, PZM, FO e TR.

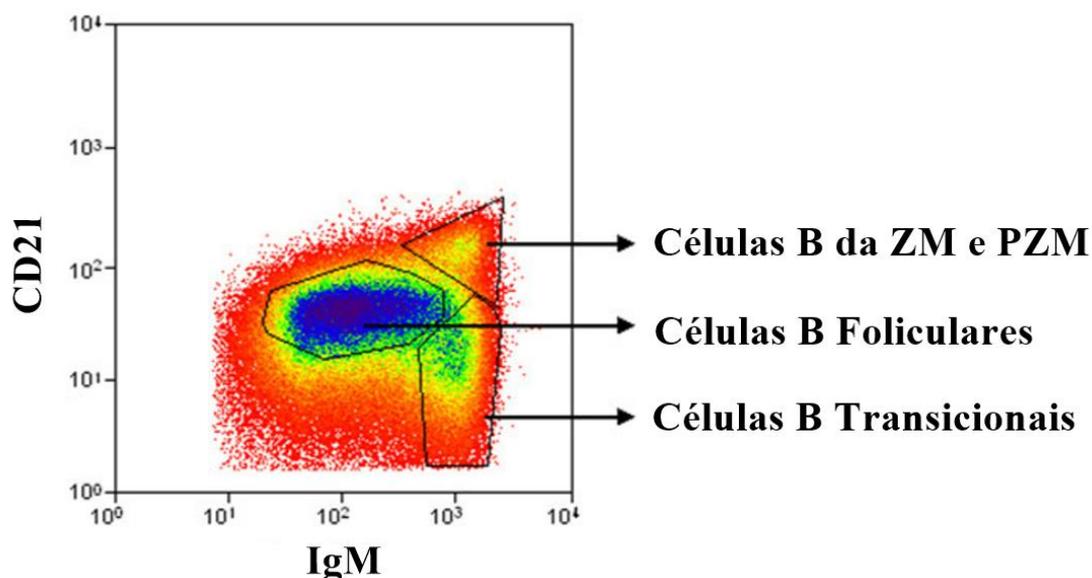


Figura 29. *Gate* padrão para classificação das populações de Zona Marginal e Pré-Zona Marginal, Foliculares e células Transicionais.

A Figura 30 (abaixo) apresenta um passo-a-passo buscando evidências de como um usuário de citometria poderia encontrar independentemente o *cluster* C2'b. A partir da vista IgM x IgD foi feito um *gate* nas células com alta expressão de IgM, mas com IgD<sup>low</sup> (Figura 30A), excluindo-se o *cluster* C5' (representado pelas linhas de contorno). Na Figura 30B (central) foi realizado outro *gate* excluindo células CD24+. Este *gate*, quando visto no plano formado por CD23 x CD21, denota uma população análoga ao *cluster* C2', contendo a ZM e uma região de pontos não pertencentes à ZM (Figura 30C). Esta população pode ser claramente subdividida em outros dois *clusters*, correspondendo a ZM e a C2'b.

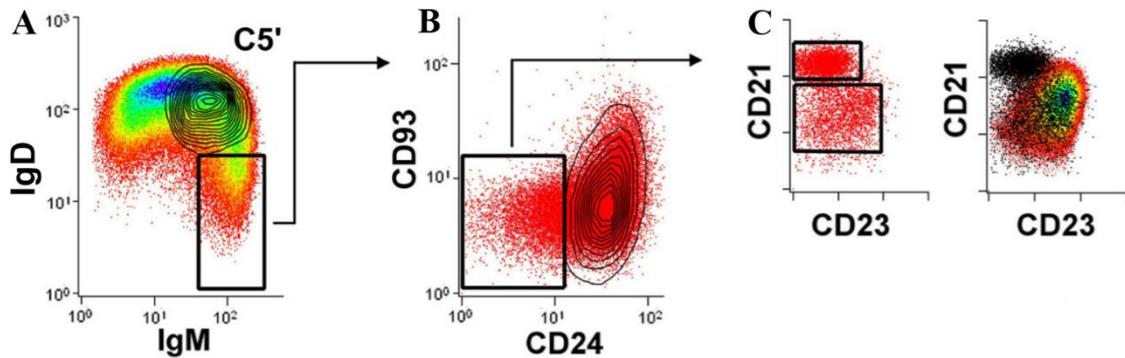


Figura 30. A Figura A apresenta a vista IgM x IgD com destaque para o *gate* em células com a configuração IgM<sup>high</sup> e IgD<sup>low</sup>. A Figura B apresenta as células selecionadas na Figura anterior (A) nas vistas CD24 x CD93. Na Figura C (em duas Figuras menores) podem ser vistas as células que foram selecionadas no *gate* anterior (Figura B) nas vistas CD23 x CD21.

Como essas células do C2'b possuem expressões fenotípicas relacionadas às células transicionais, pré-zona marginal e da zona marginal, sugere-se que as células do *cluster* C2'b seriam células do tipo B transicionais posteriores, ou então células anteriores da Pré-Zona Marginal. Em ambos os casos as células C2'b seriam células intermediárias no caminho da diferenciação celular das células transicionais para as células B da ZM (Figura 31A) em duas possíveis situações. Ou se comportando como células imediatamente anteriores a ZM (Figura 31B), ou se comportando como precursoras da PZM clássica (Figura 31C).

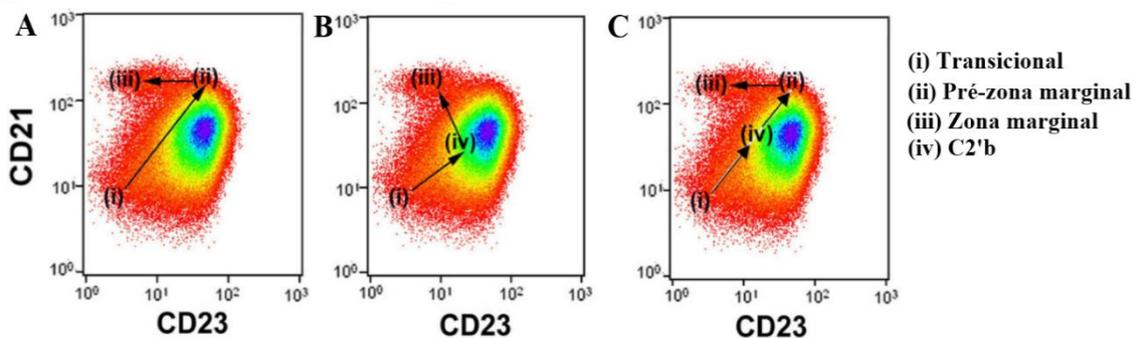


Figura 31. Com a projeção nas vistas CD23 x CD21 as três imagens apresentam os possíveis comportamentos das células a partir das células B transicionais até a zona marginal. A Figura A demarca o caminho clássico dos grupos: transicionais (i), pré-zona marginal (ii) e a zona marginal (iii). A Figura B apresenta um possível caminho de maturação da transicional (i) para o *cluster* C2'b (iv) e em seguida para a zona marginal (iii). A Figura C apresenta uma segunda possibilidade partindo da região das células transicionais (i), chegando na região do *cluster* C2'b (iv) e dela se direcionando para a pré-zona marginal (ii) para então chegar na região da zona marginal (iii).

De forma curiosa, a população da PZM clássica apresentada na Figura 32 não corresponde a nenhum *cluster* obtido na Figura 19. Em vez disso, a maior parte das células do *gate* da PZM está contida nos *clusters* C5', C4' e C2b'. Isto sugere aos autores que a

forma clássica da PZM pode não ser de fato uma PZM. Em outras palavras, o comportamento da Figura 31C poderia não existir.

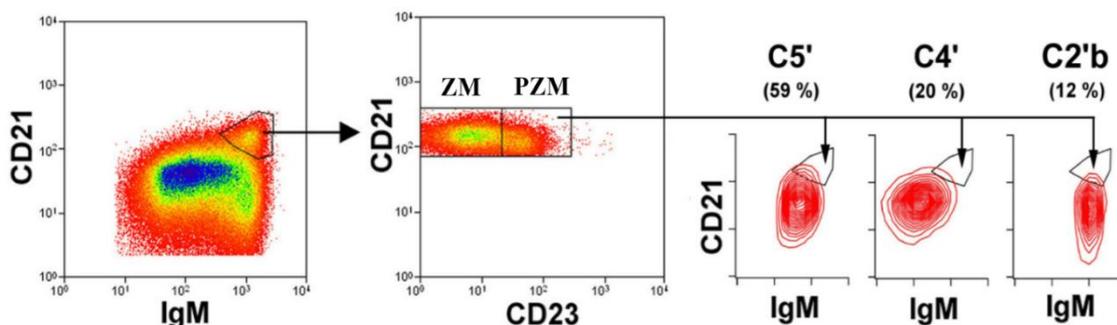


Figura 32. A Figura A apresenta a vista IgM x CD21 com um *gate* na região determinada pela zona margina e pré-zona marginal. A Figura B apresenta as células *gateadas* na Figura anterior com a separação da zona marginal (à esquerda) e da pré-zona marginal (à direita) nas vistas CD23 x CD21. A Figura C apresenta a distribuição dessas células nos *clusters* C5', C4' e C2'b nas vistas IgM x CD21, com 59%, 20% e 12% respectivamente.

Por isso, concluímos que C2'b de fato corresponde a uma subpopulação de células B e que pode ser encontrada por meio de vistas canônicas por um especialista em citometria. No entanto, para um tipo de busca não óbvio o método de Kappa modificado com o MMG pode definir um agrupamento de dados de forma semiautomática com alto grau de consistência, ainda que tenham sido utilizados centroides iniciais aleatórios no centro dos dados.

### VII.2.1 Eixo Separador

Os *clusters* gerados pelo algoritmo se apresentam de forma sobreposta para uma grande parte de suas células, contrariando o senso comum quando *gates* manuais são realizados por especialistas. Isto cria uma aparente ambiguidade na interpretação de seu resultado já que o padrão ouro é a classificação do especialista. No entanto, isto pode sugerir que as células tenham um comportamento de transformação ao longo de sua maturidade e poderiam se tornar células de um tipo ou de outro, mesmo quando observadas nas projeções canônicas clássicas CD24 x CD21 e CD23 x CD21. O fato é que para um citometrista é necessário provar que os agrupamentos realizados automaticamente pela ferramenta são, ou pelo menos deveriam ser, grupos realmente distintos, separáveis entre si. Por isso, por meio do programa IGOR foi possível realizar uma combinação linear e descobrir uma posição no espaço tridimensional, e também visível no 2-D, onde os três grandes grupos estão separáveis entre si. A Figura 24 apresentou essa condição e com isso pode-se demonstrar que os agrupamentos realizados

pela ferramenta seriam melhores quando comparados pela ação manual de um citometrista, pelo fato do algoritmo de agrupamento de dados levar em conta a multidimensionalidade.

Quando os outros dois *clusters* C2'b e C5' foram inseridos no Eixo 1 x Eixo 2, não foi possível verificar sua separabilidade em relação aos demais *clusters*. Há sobreposições claras. Também de forma empírica buscou-se encontrar, no entanto sem sucesso, uma projeção no plano que pudesse separar idealmente os cinco *clusters*. Pressupomos que não haja um plano separador para os cinco *clusters* analisados, mesmo que essa busca não tenha sido exaustiva. Sugerimos que em um trabalho futuro busque-se a utilização de ferramentas computacionais com superfícies separadoras não lineares. Caso seja possível separar estes dados, ou até outros, de forma não linear, seria mais uma conquista no aumento da confiança no uso de ferramentas computacionais junto aos citometristas.

### **VII.2.2 *Clusters* não Identificados**

Durante este trabalho foram testadas classificações entre sete e dez *clusters*. No entanto, não foi possível ao especialista em citometria, à luz do conhecimento atual, encontrar significado biológico para mais de seis agrupamentos em todos os *clusters*. Por isso, optou-se por trabalhar com seis *clusters* e garantir uma explicação biologicamente plausível sobre todos os grupos encontrados.

Nas três Figuras seguintes (33, 34 e 35) estão apresentadas uma das classificações com 10 *clusters* para as principais vistas de estudo e análise de um especialista. Nas legendas das Figuras estão descritos quais *clusters* puderam ser identificados, ou não, pelo especialista.

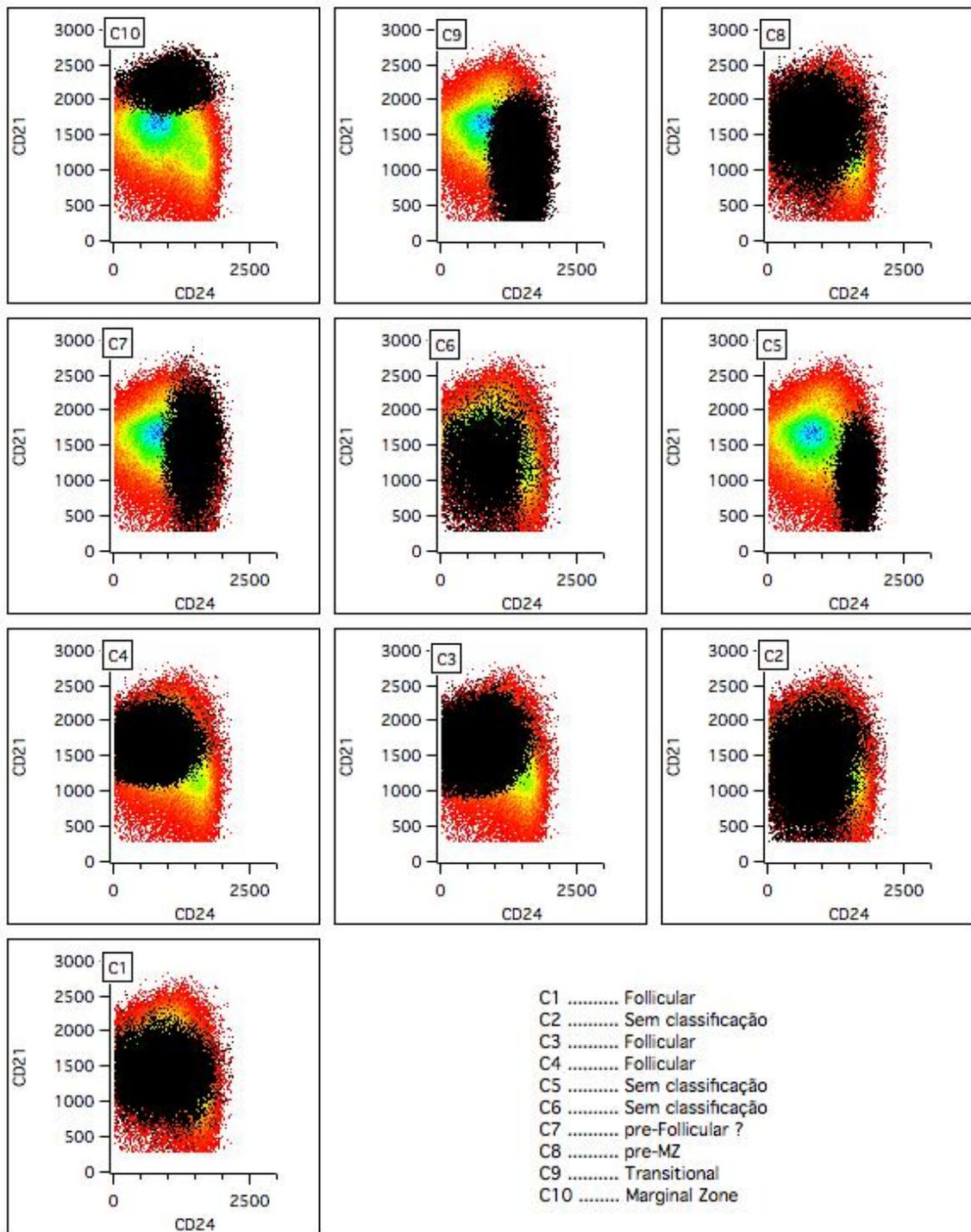


Figura 33. Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista CD24 x CD21.

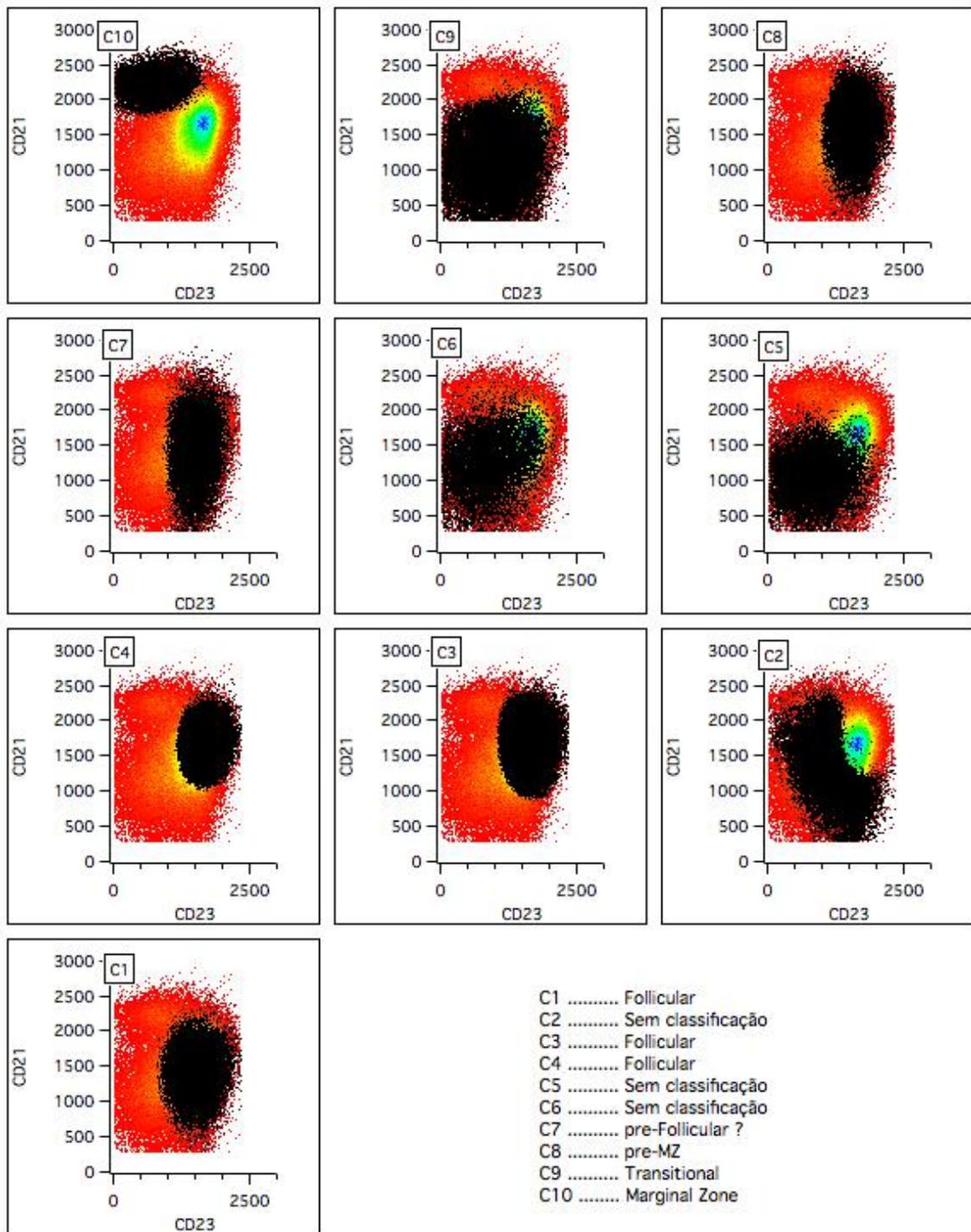


Figura 34. Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista CD23 x CD21.

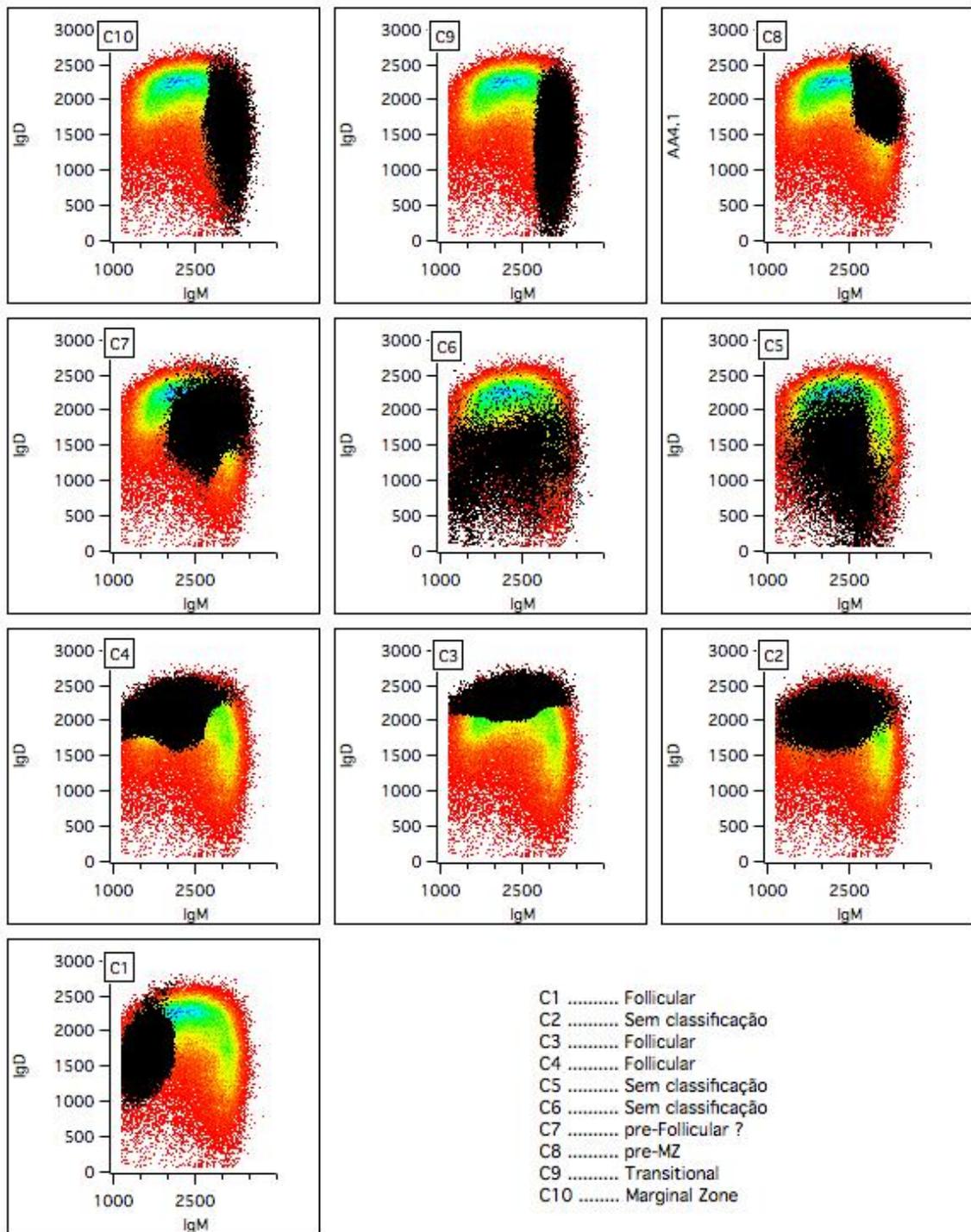


Figura 35. Apresentação de dez *clusters* (em pontos pretos) de uma classificação de dados para a vista IgM x IgD.

# CAPÍTULO 8

## CONCLUSÃO

O objetivo de desenvolver métodos de *clusterização* de dados multidimensionais aplicáveis à área de citometria de fluxo foi alcançado. Esse trabalho apresentou uma nova técnica de inicialização de centroides baseado em modas, que em junto com o BIC e o MMG deram origem a um novo método de agrupamento de dados multidimensionais que apresentou bons resultados em dados simulados e em dados de citometria de fluxo. Uma limitação do método de agrupamento de dados com inicialização de centroides determinísticos baseado em modas é sua possibilidade de não encontrar múltiplas modas caso haja uma alta concentração de dados. Quando há uma alta densidade de dados, sua projeção nos eixos das componentes principais pode se apresentar como uma distribuição unimodal, onde existe apenas um único pico (uma moda), sem nenhum outro pico aparente. Esta falta de “variação de dados” nas projeções das componentes principais compromete o uso desta técnica porque invalida o emprego de múltiplas modas para criação dos centroides. Desta forma, ficou claro que este método não é tão eficiente para investigar agrupamentos de baixa densidade que estejam próximo de agrupamentos com grande quantidade de dados (eventos raros, por exemplo).

Adicionalmente, o método de agrupamento de dados com inicialização de centroides determinísticos baseado em modas também apresenta limitações quando os agrupamentos de dados se distribuem de forma concêntrica (esférica ou circular), como, por exemplo, os dados do conjunto simulado Compound (Seção VI.1.2).

Foi também desenvolvido um novo método de agrupamento de dados quasi-determinístico baseado no coeficiente individual Kappa (CIK), chamado de método Kappa modificado, que foi útil para sugerir um novo grupo biológico relevante na área de imunologia. Apesar deste trabalho ter sugerido um novo grupo biológico no que se refere à maturação das células B, não foi possível se determinar se ele é realmente a

chamada pré-zona marginal. Para isto, será necessário um teste *in vitro* com camundongos e com o uso de um aparelho de citometria de fluxo dotado de um *sorter*, ferramenta adicionada aos citômetros de fluxo que permite a seleção individual de células. Com a inserção de determinados parâmetros previamente inseridos pelo usuário as células podem ser escolhidas, ou descartadas, pelo *sorter*, após sua etapa de medição. Tal experimentação figura-se como proposta de continuidade natural do presente trabalho.

Outra sugestão de continuidade é a utilização do método desenvolvido (método Kappa modificado) com mais de seis *clusters*. O quantitativo de seis *clusters* sugerido pelo especialista abrange somente agrupamentos de células razoavelmente estabelecidos na literatura. A investigação acima de seis agrupamentos poderia sugerir subgrupos ainda não estudados pelos métodos classicamente utilizados em citometria de fluxo, abrindo-se novas perspectivas de pesquisa.

## CONTRIBUIÇÕES CIENTÍFICAS RESULTANTES DESTE TRABALHO

- 1) Alves, E.T.A., Souza, M.N., “Ferramenta computacional para avaliação de dados de citometria de fluxo multiparamétricos”, In: **Anais do XXII Congresso Brasileiro de Engenharia Biomédica**, pp. 991-994, Nov. 2010 (ISSN: 2179-3220)
- 2) Alves, E.T.A., Frucht, L.C., Souza, M.N., Nobrega, A.F.A., “Iterative optimization algorithm - an alternative clustering tool for biological analysis using flow cytometry data”, In: **Proceedings of Pan American Health Care Exchanges (PAHCE) 2012**, pp.66-70, Mar.2012 (ISBN: 978-1-4673-1134-2).
- 3) Alves, E.T.A., Granato, A.P., Souza, M.N., Nobrega, A.F.A., “Flow Cytometry Characterization of B Cell Subsets with Multidimensional Data Analysis”, submetido à **Cytometry Part A**, Jun. 2015.
- 4) Alves, E.T.A., Nobrega, A.F.A., Souza, M.N., “Initialization of multidimensional centroids based on modes with BIC support”, submetido à **Advanced in Data Analysis and Classification**, Jul. 2015.

## REFERÊNCIAS BIBLIOGRÁFICAS

- AIKES J.J., LEE, H.D., FERRERO, C.A., *et al.*, 2012, *Estudo da influência de diversas medidas de similaridade na previsão de séries temporais utilizando o algoritmo kNN-TSP*, M.Sc. dissertação, Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos, Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, PR.
- AL-HADDAD, L., MORRIS, C., BODDY, L., "Training radial basis function neural networks: effects of training set size and imbalanced training sets". **Journal of Microbiological Methods** v. 43, n. 1, pp. 33-44, Dec. 2000.
- ALLMAN, D., PILLAI, S., "Peripheral B cell subsets", **Current Opinion Immunology** v. 20, n. 2, pp. 149-157, Aug. 2008.
- ALLMAN, D., SRIVASTAVA, B., LINDSLEY, R.C., " Alternative routes to maturity: branch points and pathways for generating follicular and marginal zone B cells", **Immunology Review**, v. 197, pp.146-160, Feb. 2004.
- ARMITAGE, J., "Bone-Marrow Transplantation", **New England Journal of Medicine** v. 330, n. 12, pp. 827-838, Mar.1994.
- ATTNEAVE, F.F., "Dimensions of similarity", **The American Journal of Psychology** v. 63, n. 4, pp. 516-556, Oct. 1950.
- BACAL, N., GUERRA, J.C.C., FERREIRA, E., *et al.*, "A importância da citometria de fluxo no diagnóstico raro de mieloma mielomonocítico", **Jornal Brasileiro de Patologia e Medicina Laboratorial** v. 38, n. 1, pp. 25-31, Out. 2002.
- BAI, L.; LIANG, J.; DANG, C., "An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data", **Knowledge-Based Systems**, v.24, n. 6, Elsevier, pp. 785-795, Aug. 2011.
- BAKKER SCHUT, T.C., GROOTH DE, B.G., GREVE, J., "Cluster Analysis of Flow Cytometric List Mode Data on a Personal Computer", **Cytometry Part A** v. 14, n. 6, Wiley-Liss, pp. 649-659. 1993.

- BARBARA, R., "Mixtures of Gaussians", **A Tutorial for the Course Computational Intelligence**, 2009. Disponível em: <  
<http://www.igi.tugraz.at/lehre/CI/SS08/tutorials/MixtGaussian/MixtGaussian.pdf>  
> Acesso em: 20 dez. 2013.
- BERTHO, A.L., **Citometria de Fluxo**. História da Citometria de Fluxo, FioCruz, Apostila. Disponível em: <[picf.ioc.fiocruz.br/apostila.doc](http://picf.ioc.fiocruz.br/apostila.doc)>. Acesso em: fev de 2013. Rio de Janeiro, 2001.
- BERTHO, A.L., SANTIAGO, M.A., DA CRUZ, A.M., *et al.*, "Detection of early apoptosis and cell death in T CD4(+) and CD8(+) cells from lesions of patients with localized cutaneous leishmaniasis", **Brazilian Journal of Medical and Biological Research**, v. 33, n. 3, pp. 317-325, Mar. 2000.
- BISHOP, C.M., **Pattern Recognition and Machine Learning**. Springer, 2006. 738 ISBN 978-0387310732.
- BODDY, L., MORRIS, C.W., WILKINS, M.F., *et al.*, "Neural-Network Analysis of Flow Cytometry Data for 40 Marine Phytoplankton Species", **Cytometry** v. 15, n. 4, pp. 283-293, Apr. 1994.
- BODDY, L., WILKINS, M.F., MORRIS, C.W., "Pattern recognition in flow cytometry", **Cytometry**, v. 44, n. 3, pp. 195-209, Jul. 2001.
- BOTEV, Z.I., GROTOWSKI, J.F., KROESE, D.P., "Kernel Density Estimation Via Diffusion", **Annals of Statistics** v. 38, n. 5, pp. 2916-2957, Aug. 2010.
- CAVALCANTI JÚNIOR, G.B., SCHEINER, M.A.M., VASCONCELOS, F.C.D., *et al.*, "Detecção da proteína p53 em células leucêmicas por citometria de fluxo e Western blot", **Revista Brasileira de Cancerologia** v. 50, n. 2, pp. 191-202, Set. 2004.
- CHAN, C., FENG, F., OTTINGER, J., FOSTER, D., WEST, M., KEPLER, T.B., "Statistical Mixture Modeling for Cell Subtype Identification in Flow Cytometry", **Cytometry Part A**, v. 83, n. 8, pp. 693-701, Aug. 2008.
- CHEN, X., HASAN, M., LIBRI, V., *et al.*, "Automated flow cytometric analysis across large numbers of samples and cell type", **Clinical Immunology**, v., 157, n. 2, Elsevier, pp. 249-260, Apr. 2015.

- COHEN, J., "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit", **Psychological Bulletin** v. 70, n. 4, pp. 213-220, Oct 1968.
- COSTA, E.S., ARROYO, M.E., PEDREIRA, C.E., *et al.*, "A new automated flow cytometry data analysis approach for the diagnostic screening of neoplastic B-cell disorders in peripheral blood samples with absolute lymphocytosis. **Leukemia** v. 20, n. 7, pp. 1221-1230, Jul. 2006.
- CROSLAND-TAYLOR, P.J., "A Device for Counting Small Particles Suspended in a Through a Tube", **Nature** v. 171, n. 4340, pp. 37-38, Jan. 1953.
- DABDOUB, S.M., RAY, WC., JUSTICE, S.S., "FIND: A new software tool and development platform for enhanced multicolor flow analysis", **BMC Bioinformatics** v. 12, n. 145, May 2011.
- DANTAS, C.A.B. **Probabilidade: Um Curso Introdutório**. 3 ed. São Paulo, EDUSP - Editora da Universidade de São Paulo, 1997. ISBN 85-314-0399-5.
- DEMPSTER, A.P., LAIRD, N.M. RUBIN, D.B., "Maximum Likelihood from incomplete data via EM algorithm", **Journal of the Royal Statistical Society Series B-Methodological** v. 39, n. 1, pp. 1-38, 1977.
- DUDA, R.O., HART, P.E., STORK, D.G., **Pattern Classification**. 2-D. New York, Wiley-Interscience, 2000. ISBN 978-0471056690.
- EISEN, D., ESSELL, J., BROUN, E.R., "Oral cavity complications of bone marrow transplantation", **Seminars in Cutaneous Medicine and Surgery**, v. 16, n. 4, p. 265-272, Dec. 1997.
- FLEISS, J.L., "Measuring nominal scale agreement among many raters", **Psychological Bulletin** v. 76, n. 5, pp. 378-382, Nov. 1971.
- FONSECA, R.J.R.M., SILVA, P.J.S.P., SILVA, R.R., "Acordo inter-juízes: O caso do coeficiente kappa", **Laboratório de Psicologia**, v. 5, n. 1, pp. 81-90, 2007.
- FRANCA, C.M., DOMINGUES-MARTINS, M., VOLPE, A., *et al.*, "Severe oral manifestations of chronic graft-vs.-host disease", **Journal of the American Dental Association** v. 132, n. 8, pp. 1124-1127, Aug. 2001.

- FRELINGER, J., OTTINGER, J., GOUTTEFANGEAS, C., *et al.*, "Modeling flow cytometry data for cancer vaccine immune monitoring", **Cancer Immunology Immunotherapy**, v. 59, n. 9, pp. 1435-1441, Sep. 2010.
- GASSEN, S.V., CALLEBAUT, B., HELDEN, M.J.V., LAMBRECHT, B.N., DEMEESTER, P., DHAENE, T., SAEYS, Y., " FlowSOM: Using self organizing maps for visualization and interpretation of cytometry data", **Cytometry Part A**, v. 87, n. 7, pp.636-645, Jan. 2015.
- GEJGUS, P., PLACEK, J., SPERKA, M., "Skin color segmentation method based on mixture of gaussians and its application in learning system for finger alphabet". In: **Proceedings of the 5th International Conference on Computer systems and technologies**, pp. 1-6, New York, USA, 2004.
- GODAVARTI, M., RODRIGUEZ J.J., YOPP, T.A., *et al.*, "Automated particle classification based on digital acquisition and analysis of flow cytometric pulse waveforms", **Cytometry** v. 24, n. 4, pp. 330-339, Aug. 1996.
- GRANATO, A., HAYASHI E.A., BAPTISTA, B.J., *et al.*, "IL-4 regulates Bim expression and promotes B cell maturation in synergy with BAFF conferring resistance to cell death at negative selection checkpoints", **Journal of Immunology** v. 192, n. 12, pp. 5761-5775, Jun. 2014.
- GRATAMA, J.W., D'HAUTCOURT, J.L, MANDY, F., *et al.*, "Flow cytometric quantitation of immunofluorescence intensity: Problems and perspectives. European Working Group on Clinical Cell Analysis", **Cytometry** v. 33, n. 2, pp. 166-178, Oct. 1998.
- GUINAMARD, R., OKIGAKI, M., SCHLESSINGER, J., *et al.*, "Absence of marginal zone B cells in Pyk-2-deficient mice defines their role in the humoral response", **Nature Immunology** v. 1, n. 1, pp. 31–36, Jul. 2000.
- HASAN, M., BEITZ, B., ROUILLY, V., *et al.*, " Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping", **Clinical Immunology**, v.157, n. 2, pp. 261-276,
- HERZENBERG, L.A., TUNG, L., MOORE, W.A., *et al.*, "Interpreting flow cytometry data: a guide for the perplexed", **Nature Immunology** v. 7, n. 7, pp. 681-686, Jul. 2006.

- JONKER, R., GROBEN, R., TARRAN, G., *et al.*, "Automated identification and characterisation of microbial populations using flow cytometry: the AIMS project", **Scientia Marina** v. 64, n. 2, pp. 225-234, Jun. 2000.
- JUNQUEIRA, L.C.; CARNEIRO, J., **Histologia Básica**. 10 ed. Rio de Janeiro, Editora Guanabara Koogan S.A., 2004.
- KAINULAINEN, J.J., **Clustering Algorithms: Basics and Visualization**, Laboratory of Computer and Information Science, Helsinki University of Technology, p. 29, Jan. 2002.
- KINDT, T.J., GOLDSBY, R.A., OSBORNE, B.A., **Imunologia de Kuby**. 6<sup>a</sup> ed. Porto Alegre, Artmed, 2008. ISBN-10: 8536311916
- KITSOS, C.M., BHAMIDIPATI, P., MELNIKOVA, I., *et al.*, "Combination of automated high throughput platforms, flow cytometry, and hierarchical clustering to detect cell state", **Cytometry Part a** v. 71A, n. 1, pp. 16-27, Jan. 2007.
- KOTSIANTIS, S.B., "Supervised Machine Learning: A Review of Classification Techniques", **Informatica** v. 31, pp. 249-268, Jul. 2007.
- KUMARARATNE, D.S., BAZIN, H., and MACLENNAN, I.C., "Marginal zones: the major B cell compartment of rat spleens". **European Journal of Immunology** v.11, n. 11, pp.858–864, Nov. 1981.
- KUMARARATNE, D.S., MACLENNAN, I.C. "The origin of marginal-zone cells", **Advances in Experimental Medicine and Biology**, v. 149, pp. 83–90, 1982.
- LANDIS, J.R., KOCH, G.G., "The measurement of observer agreement for categorical data", **Biometrics** v. 33, n. 1, pp. 159-174, Mar. 1977.
- LE MEUR, N., ROSSINI, A., GASPARETTO, M., *et al.*, "Data quality assessment of ungated flow cytometry data in high throughput experiments", **Cytometry Part a** v. 71A, n. 6, pp. 393-403, Jun. 2007.
- LIANG, J., ZHAO, X., LI, D., *et al.*, " Determining the number of clusters using information entropy for mixed data", **Pattern Recognition**, v. 45, n. 6, Elsevier, pp. 2251-2265, Jun. 2012.

- LINDEN, R., "Técnicas de Agrupamento", **Revista de Sistema de Informação da FSMA**, v. 4, pp. 18-36, Dec. 2009.
- LIU, Y.J., OLDFIELD, M., MACLENNAN, I.C.M., "Memory B cell in T cell-dependent antibody responses colonize the splenic marginal zones", *European Journal Immunology* v. 18, n. 3, pp. 355–362, Mar. 1988.
- LIZARD, G., "Flow Cytometry analyses and bioinformatics: Interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research", **Cytometry Part a** v. 71A, n. 9, pp. 646-647, Sep 2007.
- LO, K., BRINKMAN, R.R., GOTTARDO, R., "Automated Gating of flow cytometry data via robust model-based clustering". **Cytometry Part a** v. 73, n. 4, pp. 321-332, Apr. 2008.
- LO, K., HAHNE, F., BRINKMAN, R.R., *et al.*, "flowClust: a Bioconductor package for automated gating of flow cytometry data", **BMC Bioinformatics** v. 10, n. 145, May 2009.
- LU, Y., SAKAMAKI, S., KURODA, H., *et al.*, "Prevention of lethal acute graft-versus-host disease in mice by oral administration of T helper 1 inhibitor, TAK-603", **Blood** v. 97, n. 4, pp. 1123-1130, Feb. 2001.
- MALEK M., TAGHIYAR, M.J., CHONG, L., *et al.*, " flowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification", **Bioinformatics**, v. 31, n. 4, Oxford University Press, pp. 606-607, Oct. 2014.
- MANN, P.S., **Introductory Statistics**. 2nd. John Wiley & Sons, 1995. ISBN 0-471-31009-3
- MARTIN, F., KEARNEY, J.F., "Marginal-zone B cells", **Nature Reviews Immunology** v. 2, n. 5, pp. 323–335, May 2002.
- MARTINEZ, W., MARTINEZ, A., **Computational Statistics Handbook with MATLAB**. Boca Raton, Florida: Chapman & Hall/CRC, 2008. ISBN 978-1-4200-3563-6

- MASSUMOTO, C., MIZUKAMI, S., "Transplante autólogo de medula óssea e imunoterapia pós-transplante", **Medicina Ribeirão Preto** v. 33, n. 4, pp. 405-414. Out. 2000.
- MESKAS, J., DROUMEVA, R., "flowCL: Semantic labelling of flow cytometric cell populations", **Bioconductor**, p. 11, April, 2015. Disponível em: <<http://www.bioconductor.org/packages/release/bioc/vignettes/flowCL/inst/doc/flowCL.pdf>>. Acesso em: 20 de maio de 2015.
- McHUGH, M.L., "Interrater reliability: the kappa statistic". **Biochemia Medica** v. 22, n.3, pp. 276-282, 2012.
- McLACHLAN, G.J., DAVID, P., **Finite Mixture Models**. 1 ed. New York: Wiley-Interscience, 2000.
- McLACHLAN, G. J., KRISHNAN, T., **The EM Algorithm and Extensions**. 2 ed. New Jersey: Wiley, 2007. ISBN-13: 978-0471201700
- MEDHI, J., **Statistical Methods: An Introductory Text**. New Delhi: New Age International (P) Limited, 1992. 441 ISBN 81-224-0419-7.
- MODESTO, T.M., NEVES, M.A.B., DE BRITO, A.E., *et al.*, "Importância e Vantagem da Citometria de Fluxo Frente Aos Testes De Triagem No Diagnóstico Da Hemoglobinúria Paroxística Noturna", **Rev. Bras. Hematologia e Hemoterapia** v. 28, n. 4, pp. 275-279, Out. 2006.
- MOON, T.K., "The expectation-maximization algorithm", **IEEE Signal Processing Magazine** v. 13, n. 6, pp. 47-60, Nov. 1996.
- MORRIS, C.W., AUTRET, A., BODDY, L., "Support vector machines for identifying organisms - a comparison with strongly partitioned radial basis function networks", **Ecological Modelling** v. 146, n. 1-3, pp. 57-67, Dec. 2001.
- NAKAGE, A.P.M., SANTANA, A.E., DE CAPUA, M.L.B, *et al.*, "Metodologia e aplicação da citometria de fluxo na hematologia veterinária", **Ciência Rural**. v. 35, n. 4, pp. 966-973, Jul. 2005.
- NUNES, B. P., 2006, *Classificação automática de dados semi-estruturados*. M.Sc. dissertação, Programa de Pós-Graduação em Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ.

- ORMEROD, M.G., **Flow Cytometry: A Practical Approach**. 3 ed. Texas, Oxford University Press, 2000. ISBN 0199638241, 9780199638246.
- PAKHIRA, M.K., "Finding Number of Clusters before Finding Clusters", **Procedia Technology**, v. 4, Elsevier, pp. 27-37, Feb. 2012.
- PANGBORN, A. D., 2009, *Accelerating Flow Cytometry Data Clustering Workflows with Graphics Processing Units*. M.Sc. dissertation, Computer Engineering, Rochester Institute of Technology, Rochester.
- PARKS, D.R., HERZENBERG, L.A., "Fluorescence-activated cell sorting: Theory, experimental optimization and applications in Lymphoid cell biology", **Methods Enzymol** v. 108, pp. 197-224 I, 1984.
- PEEL, D., McLACHLAN, G.J., "Robust mixture modelling using the t distribution", **Statistic and Computing** v. 10, pp. 339-348. 2000.
- PICARD, F. **An Introduction to Mixture Models**. In: Research Report N.7, Statistic for Systems Biology Group, Paris, France, 2007.
- POWERS, D.M.W., "The Problem with Kappa", In: **Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics**, Centre for Knowledge & Interaction Technology, CSEM, pp. 345–355, Avignon, France, Apr. 2012.
- QIU, P., "Computational Prediction of Manually Gated Rare Cells in Flow Cytometry Data", **Cytometry Part A**, v. 87, n. 7, pp. 594-602, Mar. 2015.
- QUINN, J. *et al.*, "A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow", **Cytometry Part a** v. 71A, n. 8, pp. 612-624, Aug. 2007.
- RATCOM, INC., THOMAS, R.A., EGGLESTON, R.W., **Flow Cytometry**. US4673288 A, SET. 1984.
- SCHWARZ, G., "Estimating the dimension of a model". **The Annals of Statistics**, v. 6, n. 2, pp. 461-464, Mar. 1978.

- SHAPIRO, H.M., **Practical Flow Cytometry**. 4 ed, John Wiley & Sons, 2005. 736  
ISBN: 0471434035, ISBN 13: 9780471434030.
- SILVA, D., SANTOS, V., *ClusteringTools: Uma Ferramenta de Auxílio ao Ensino de Técnicas de Clusterização*, Monografia, Departamento de Ciência da Computação, Instituto de Matemática da Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 2007.
- SIM, J., WRIGHT, C.C., "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements", **Physical Therapy** v. 85, n. 3, pp. 257-268, Mar. 2005.
- SMITH, P.J., KHAN, I.A., ERRINGTON, R.J., "Cytomics and drug development", **Cytometry Part a** v. 71A, n. 6, pp. 349-351, Jun. 2007.
- SRIVASTAVA, B. QUINN III, W.J. HAZARD, H., *et al.*, "Characterization of marginal zone B cell precursor", **The Journal of Experimental Medicine** v. 202, n. 9, pp. 1225-1234. Oct. 2005.
- SVENSSON, C.M., KRUSEKOPF, S., LÜCKE, J., *et at.*, "Automated Detection of Circulating Tumor Cells with Naive Bayesian Classifiers", **Cytometry Part A**. v.85A, n. 6, pp. 501-511, Jun. 2014.
- VEZHNEVETS, V., SAZONOV, V., ANDREEVA, A., "A Survey on Pixel-Based Skin Color Detection Techniques". In: **Proceedings of the GraphiCon**, pp. 85-92, Moscow, Russia, Sep. 2003.
- VIEIRA, S. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro, Elsevier Editora, 2008.  
ISBN 978-85-352-2843-4.
- WANDERLEY, H., 2006. *Construção e Validação de um Citômetro de Fluxo Monoparamétrico*. M.Sc. dissertação, Programa de Engenharia Biomédica, COPPE/UFRJ, Rio de Janeiro, RJ.
- WEINBERG, S.L., ABRAMOWITZ, S.K., **Statistics Using SPSS: An Integrative Approach**. 2 ed. Cambridge University Press, 2008. 764, ISBN: 0521899222, ISBN 13: 9780521899222

- WELLER, A.F., HARRIS, A.J., WARE, J.A., "Two supervised neural networks for classification of sedimentary organic matter images from palynological preparations", **Mathematical Geology** v. 39, n. 7, pp. 657-671, Oct. 2007.
- WILKINS, M., BODDY, L., MORRIS, C.W., *et al.*, "Identification of phytoplankton from flow cytometry data by using radial basis function neural networks", **Applied and Environmental Microbiology** v. 65, n. 10, pp. 4404-4410, Oct. 1999.
- WILKINS, M., HARDY, S.A., BODDY, L., *et al.*, "Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data", **Cytometry**, v. 44, n. 3, pp. 210-217, Jul. 2001.
- WUNSCH, D., XU, R., "Survey of clustering algorithms", **IEEE Transactions on Neural Networks** v. 16, n. 3, pp. 645-678, May 2005.
- YU, H., LIU, Z., WANG, G., "An automatic method to determine the number of clusters using decision-theoretic rough set", **International Journal of Approximate Reasoning**, v. 55, n. 1, Elsevier, pp. 101-115, Jan. 2014.
- ZADEH, L. A. "Fuzzy sets", **Information and Control**, n. 8, pp. 338-353, 1965.
- ZARE, H., SHOOSHTARI, P., GUPTA, A., *et al.*, "Data reduction for spectral clustering to analyze high throughput flow cytometry data", **BMC Bioinformatics** v. 11, n. 1, p. 403, Jul. 2010.
- ZARE, H., SHOOSHTARI, P., **SamSPECTRAL: A Modified Spectral Clustering Method for Clustering Flow Cytometry Data.** pp. 1-11, Oct. 2010. Disponível em: <  
[http://www.bioconductor.org/packages/release/bioc/vignettes/SamSPECTRAL/inst/doc/Clustering\\_by\\_SamSPECTRAL.pdf](http://www.bioconductor.org/packages/release/bioc/vignettes/SamSPECTRAL/inst/doc/Clustering_by_SamSPECTRAL.pdf)>. Acesso em: 22 de abril de 2012.
- ZENG, Q.T., PRATT, J.P., PAK, J., *et al.*, "Feature-guided clustering of multi-dimensional flow cytometry datasets", **Journal of Biomedical Informatics** v. 40, n. 3, pp. 325-331, Jun. 2007.

## APÊNDICE A – ALGUMAS DEFINIÇÕES MATEMÁTICAS

Este apêndice apresenta alguns conceitos que foram aplicados no desenvolvimento deste trabalho.

### 1 *Cluster*, Agrupamento de Dados, Classificação e Centroide

Para efeito desta escrita um *cluster* será o conjunto numérico de dados que pode ser representado graficamente e que seus elementos possuem certo grau de similaridade entre si. Em citometria de fluxo é comum o emprego de cores para demarcar *clusters*. A Figura 1 apresenta um exemplo onde podem ser visualizados facilmente três *clusters*, marcados pelas cores azul, verde e vermelho. A demarcação dessas cores significa que dados com a mesma cor são similares em relação ao conjunto observado e por isso estes são do mesmo *cluster*.

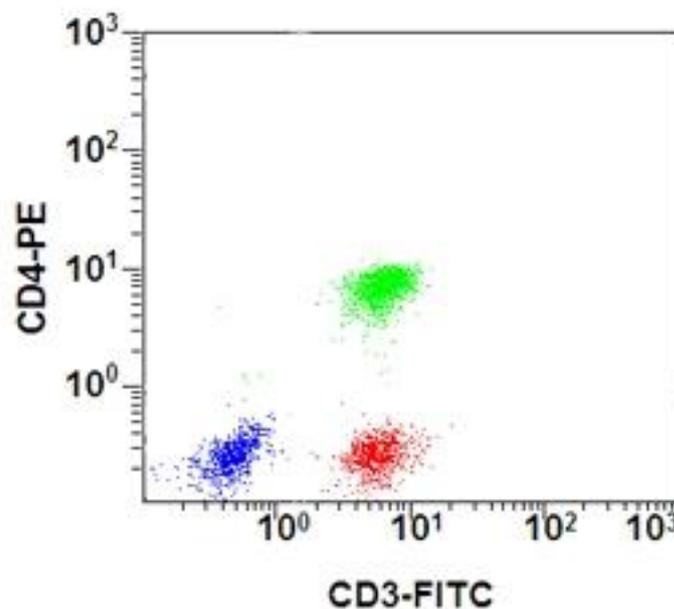


Figura 1. Exemplo de agrupamento de dados em citometria. Os eixos representam a intensidade luminosa no comprimento de onda de cada um dos fluorocromos utilizados na coleta (extraído e adaptado de [http://www.appliedcytometry.com/flow\\_cytometry.php](http://www.appliedcytometry.com/flow_cytometry.php)).

As definições a seguir são aplicáveis a quaisquer *clusters* (KAINULAINEN, 2002). No entanto, existem exceções quando aplicado o conceito de lógica fuzzy, que considera que uma amostra pode pertencer a mais de um *cluster* com níveis de probabilidade diferentes. As definições de KAINULAINEN (2002) são:

- 1) Um *cluster* não pode ser um conjunto vazio;
- 2) Cada amostra pertence a um *cluster*, e;
- 3) Cada amostra pertence a um único *cluster*.

Apesar das definições de KAINULAINEN (2002) serem bem claras, percebe-se que a terceira definição engloba a segunda definição.

O termo agrupamento de dados irá se referir à capacidade de um sistema organizar os dados em *clusters*, ou seja, é o processo de agrupar esses dados por algum critério de semelhança. Já classificação refere-se à capacidade de identificar um elemento e atribuí-lo a um grupo previamente conhecido (SILVA e SANTOS, 2007). Obviamente o melhor agrupamento possível é aquele que pode analisar todas as possibilidades e escolher o melhor, segundo um critério já definido. Entretanto, isso nem sempre é viável, devido a fatores como, por exemplo, o custo computacional (KAINULAINEN, 2002).

Já o ponto central de *cluster* é um centroide. Ele é a representação gráfica, e numérica, do valor médio de um conjunto de dados dependendo da métrica utilizada para se calcular a similaridade entre os vários dados que o constituem. Devido ao fato da citometria de fluxo utilizar vários canais de fluorescência, dados citométricos são considerados multivariados, sendo a dimensão do centroide de um *cluster* de dados citométricos igual ao número de canais dos dados.

## 2 Medidas de Similaridade

Segundo AIKES (2012) a medição de similaridade define o critério para quantificar quão similares são duas sequências; e decidir se estas serão tidas como pertencentes a um determinado grupo, ou não. As métricas mais expressivas derivadas da norma são: distância Euclidiana, Manhattan e Mahalanobis.

A distância Euclidiana é a mais conhecida das medidas de similaridade (eq. 1)

$$d_{xy} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

onde,  $x$  e  $y$  representam os vetores dos dados,  $i$  o índice que se refere a cada componente do vetor e  $n$  é a dimensão.

A distância de Manhattan, ou *City Block*, (eq. 2) é de certo modo similar à distância Euclidiana, com a diferença se usar a norma da diferença entre cada componente, ao invés da diferença ao quadrado. Isto provoca um menor impacto na distância quando a diferença entre os valores das componentes é grande, já que este não é elevado ao quadrado.

$$d_{xy} = \sum_{i=1}^n |x_i - y_i| \quad (2)$$

Por fim, a distância de Mahalanobis se diferencia da distância Euclidiana por levar em conta a dispersão do conjunto de dados (LINDEN, 2009). Esta métrica é bastante útil para se avaliar medidas de distâncias de conjuntos de dados com dispersões diferentes, uma vez que ela considera a dispersão na determinação ou normalização da medida de distância. Formalmente, a distância de Mahalanobis entre dois dados multidimensionais  $\mathbf{x}$  e  $\mathbf{y}$ , pertencentes a um conjunto de dados com matriz de covariância  $\Sigma$  pode ser expressa pela equação 3.

$$d_{xy} = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})} \quad (3)$$

É fácil verificar que se a matriz de covariância for igual à matriz identidade, a distância de Mahalanobis torna-se igual à distância Euclidiana. Além disso, a distância de Mahalanobis entre um elemento  $\mathbf{x}$  e o centroide de um *cluster*, sendo este último representado pelo vetor das médias nas várias dimensões  $\boldsymbol{\mu} = \{\mu_1, \mu_2, \dots, \mu_n\}$ , pode ser escrita como na equação 4.

$$d_x = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (4)$$

A equação 4 mede a distância de um elemento ao centro de um *cluster* com base na sua distribuição dos elementos destes *cluster*.

### 3 Máxima Verossimilhança

Suponha uma amostra aleatória  $\mathbf{x} = (x_1, x_2, x_3, \dots, x_n)$  retirada de uma população com uma probabilidade  $p(\mathbf{x}, \boldsymbol{\theta})$ , sendo  $\boldsymbol{\theta}$  um vetor com os parâmetros desconhecidos. A função de verossimilhança  $L$  é definida como:

$$L(\boldsymbol{\theta}; x_1, \dots, x_n) = p(x_1; \boldsymbol{\theta}) * p(x_2; \boldsymbol{\theta}) * \dots * p(x_n; \boldsymbol{\theta}) = \prod_{i=1}^n p(x_i; \boldsymbol{\theta}) \quad (5)$$

A estimativa por máxima verossimilhança de  $\theta$  é o valor  $\theta^*$  que maximiza  $L(\theta; x_1, \dots, x_n)$ . É possível encontrar o máximo de  $L(\theta)$  por meio da derivada de  $L$  igualada a zero (eq. 6), sendo que a solução desta leva ao valor  $\theta^*$ .

$$\frac{dL(\theta)}{d\theta} = 0 \quad (6)$$

Em muitas aplicações é mais simples algebricamente (e muitas vezes computacionalmente) trabalhar na escala dos logaritmos. Do ponto de vista da maximização não faz diferença, já que a função logaritmo é crescente e a variável  $\theta$  que maximiza  $L(\theta)$  é o mesmo que maximiza  $\ln L(\theta)$  (Ehlers, 2009). A função  $\ln L = \ln L(\theta)$  é denominada função de log-verossimilhança. Assim, nota-se que:

$$\ln(L(\theta)) = \ln(p(x_1)) + \ln(p(x_2)) + \dots + \ln(p(x_n)) = \sum_{i=1}^n \ln(p(x_i)) \quad (7)$$

Desse modo, o máximo também pode ser encontrado por:

$$\frac{d \ln L(\theta)}{d\theta} = 0 \quad (8)$$

Como, em geral, a derivada da soma é mais simples do que a derivada do produto, o logaritmo da verossimilhança é mais fácil de ser maximizado que a verossimilhança. Cada iteração desse estimador envolve dois passos. Um chamado de “E” (esperança matemática) e um passo chamado de “M” (maximização). O passo “E” calcula o valor esperado do logaritmo da função de verossimilhança. Enquanto o passo “M” procura seu valor máximo. De forma geral esses dois passos são executados ao longo do algoritmo até que seja atingida uma convergência.

## APÊNDICE B – APRENDIZADO SUPERVISIONADO E NÃO-SUPERVISIONADO

Existem atualmente diversas técnicas de aprendizado na área de inteligência artificial, as quais tentam explorar semelhanças entre padrões similares (KOTSIANTIS, 2007). Estas técnicas se dividem em dois grandes grupos: as de aprendizado supervisionado e as de aprendizado não supervisionado. A primeira técnica pressupõe que um conjunto de dados será utilizado a priori como modelo para o desenvolvimento de um sistema de classificação de um segundo conjunto de dados. Genericamente falando, o primeiro conjunto é chamado de conjunto de treinamento e o segundo de conjunto de teste. Às vezes, também é considerado um terceiro conjunto chamado de conjunto de validação, que tem por finalidade melhorar o desenvolvimento do sistema de classificação.

De forma geral, o aprendizado supervisionado pressupõe um conjunto de dados já rotulados que servem de referência para o desenvolvimento do classificador. Entende-se por dados rotulados, dados cujo *cluster* é conhecido. O desafio neste tipo de aprendizado está em otimizar os parâmetros internos do algoritmo de agrupamento para que outras entradas desconhecidas também possam ser rotuladas adequadamente. Na literatura podem ser encontrados vários algoritmos para aprendizado supervisionado. Cada um deles utiliza uma técnica diferente para determinar sua função de aprendizagem. Dependendo da aplicação, alguns algoritmos são mais recomendados que outros. Algoritmos clássicos de aprendizado supervisionado são as Árvores de Decisão, Vizinho Mais Próximo, Máquina de Vetores de Suporte, Redes Neurais Artificiais, entre outros (BODDY, WILKINS *et al.*, 2001; NUNES, 2009).

No aprendizado não supervisionado não há modelo a priori para ser seguido. o agrupamento dos dados é gerada ao longo da etapa de iteração do algoritmo. A este tipo de aprendizado é dada a tarefa de *clusterizar* os dados durante suas iterações (WUNSCH e XU, 2005). As etapas gerais de um processo de aprendizagem não supervisionada são (DE LIMA, 2011):

- a) seleção de atributos;
- b) medida de similaridade;
- c) critério de agrupamento;

- d) algoritmo de agrupamento;
- e) validação, e;
- f) interpretação dos resultados.

Os atributos, ou seja, os dados precisam já estar processados de forma à ferramenta de trabalho poder lê-los e interpretá-los. A medida de similaridade, já discutida anteriormente, deve subsidiar o critério de agrupamento na decisão de agrupar um dado em um *cluster*, ou em outro. Depois de adotados uma métrica de similaridade e um critério a ser otimizado, cabe ao algoritmo de agrupamento de dados “revelar a estrutura de agrupamento dos dados”. É importante se ter em mente que diferentes critérios, métricas ou algoritmos, fornecem resultados bastante diferentes para um mesmo conjunto de dados (KOTSIANTIS, 2007). A validação do desenvolvimento de um classificador geralmente é realizada por meio de testes apropriados, como por exemplo, inserir um conjunto de dados conhecidos e analisar sua saída. Por fim, tem-se a etapa de interpretação dos resultados. Em dados de citometria essa interpretação cabe ao especialista da área, o qual compara os resultados do agrupamento com sua experiência acumulada ou com evidências experimentais.

Um algoritmo não supervisionado bastante conhecido na literatura é o *k-means*, onde *k clusters* devem ser usados para classificar *n* pontos (ou dados). Geralmente nesse algoritmo utiliza-se como métrica a distância Euclidiana. A metodologia básica de sua operação pode ser resumida em 5 passos:

- 1) *Inicialização de centroides*. O usuário, ou o programa de computador, deve fornecer a posição inicial dos centroides, de forma aleatória ou arbitrariamente escolhida.
- 2) *Calcular distâncias*. Para cada *cluster*, calcula-se a distância entre o respectivo centroide e os *n* pontos da amostra de dados, sendo criada uma matriz de distância. Esse passo é geralmente o mais custoso computacionalmente.
- 3) *Agrupamento*. Cada ponto é agrupado no *cluster* que possui o centroide considerado mais próximo. Se não há mudança no agrupamento dos pontos em relação a uma iteração anterior, normalmente o algoritmo finaliza sua operação neste passo.
- 4) *Reposicionamento dos centroides*. Para os centroides que receberam pontos, suas coordenadas são modificadas com essas alterações. Analogamente, o mesmo acontece para os centroides que perderam pontos.

5) O algoritmo retorna ao passo 2 realizando novas medidas de distância.

Uma das desvantagens mais comentadas do *k-means* está no fato de se definir à priori o número de *clusters*. Uma escolha inadequada pode comprometer o resultado. Uma variante desse algoritmo é o método desenvolvido por Bakker Schut e colaboradores (BAKKER SCHUT, GROOTH DE *et al.*, 1993). Os autores propuseram que o algoritmo *k-means* iniciasse com um número alto de centroides e que gradativamente esse número fosse diminuído para o número final de *clusters* previamente definido. A metodologia empregada para essa diminuição foi a de agrupar centroides que estivessem mais próximos. Um critério de validação foi estabelecido para a medição dessa distância crítica. O objetivo do trabalho foi o de diminuir algumas das limitações originais do *k-means*, como a dependência da posição inicial dos centroides (BAKKER, GROOTH DE *et al.*, 1993).

Outra característica relacionada aos algoritmos para agrupamento é aquela que diz respeito a sua total automatização, ou seja, o quanto um programa deve ser independente do usuário. É interessante notar que nem sempre uma ferramenta totalmente automática é a melhor opção. Por exemplo, quando há uma clara necessidade de interpretação humana sobre um dado biológico, ferramentas semiautomáticas tornam-se mais interessantes que as automáticas (COSTA, ARROYO *et al.*, 2006). Em várias situações uma ferramenta semiautomática é a mais indicada por poder sugerir ao especialista a quantidade de *clusters* adequada para o conjunto de dados em estudo. Cabendo apenas ao especialista acatar, ou não, as sugestões da ferramenta.

## APÊNDICE C – COEFICIENTE KAPPA

Apesar da literatura apresentar o cálculo do Coeficiente Kappa (FLEISS, 1971), este apêndice busca apresentar de forma ilustrativa como ele é calculado.

Considere exames médicos onde podem ser considerados três diagnósticos: resfriado, gripe ou pneumonia. Dois médicos (A, B) foram chamados para que dessem seus laudos. Cem exames foram avaliados. A Tabela 1 apresenta a distribuição de classificações.

Tabela 1. Exemplo para cálculo de Kappa simples sem peso.

		Médico A			Total
		Resfriado	Gripe	Pneumonia	
Médico B	Exames				
	Resfriado	21	5	3	29
	Gripe	12	21	0	33
	Pneumonia	2	10	26	38
Total	35	36	29	100	

Para efeito didático será calculado o coeficiente Kappa apenas pela equação:

$$k = \frac{\sum fa - \sum fe}{N - \sum fe} \quad (1)$$

A frequência de concordância nas mesmas categorias ( $\sum fa$ ) é obtida pela soma dos valores das células que se encontram na diagonal principal,  $21+21+26=68$ . Já para a frequência de observações feitas ao acaso ( $\sum fe$ ), é obtida pela soma dos produtos entre *total linha X total coluna* de cada categoria, dividindo-se pelo total de observações (Fonseca, Silva *et al.*, 2007). Assim, temos:

$$\sum fe = \frac{29 \cdot 35}{100} + \frac{33 \cdot 36}{100} + \frac{38 \cdot 29}{100} = 33,05$$

Utilizando a equação 1, temos que o índice Kappa será:

$$k = \frac{68 - 33,05}{100 - 33,05} = \frac{34,95}{66,95} = 0,52$$

Quando se tem mais do que 2 observadores distintos, a obtenção do coeficiente Kappa é realizada por uma variação da fórmula anterior. Novamente, de forma didática, considere o seguinte exemplo abaixo para um melhor entendimento.

Considere um exame que pode ser classificado em 3 patologias. Cinco médicos foram chamados para darem seus laudos em 6 exames. Deseja-se saber qual o grau de concordância entre os cinco médicos de acordo com as classificações apresentadas na Tabela 2.

Tabela 2. Exemplo para cálculo do coeficiente Kappa com múltiplos observadores. Em cada célula encontra-se o número de médicos que diagnosticou a patologia indicada pela coluna correspondente.

Exames	Categorias		
	A	B	C
1	5	0	0
2	1	2	2
3	0	4	1
4	0	0	5
5	1	4	0
6	2	3	0

Neste exemplo do exame 1 temos que:  $N = 6$ ,  $n = 5$  e  $t = 3$ . Para iniciar o cálculo de  $P_o$  é importante lembrar que:

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i \quad (2)$$

$$P_i = \frac{1}{n(n-1)} [(\sum_{j=1}^t x_{ij}^2) - n] \quad (3)$$

É preciso inicialmente se determinar a proporção de classificações nas mesmas categorias ( $P_i$ ), conforme a equação 3. Para o primeiro exame ( $i=1$ ), temos que:

$$P_i = \frac{1}{n(n-1)} [(\sum_{j=1}^t x_{ij}^2) - n] = \frac{1}{5(5-1)} [(5^2 + 5^2 + 0^2) - 5] = \frac{20}{20} = 1$$

Calculando-se para os demais exames a mesma regra, temos o seguinte resultado na Tabela 3.

Tabela 3. Exemplo para cálculo do coeficiente Kappa para múltiplos observadores com  $P_i$  atualizado.

Exames	Categorias			$P_i$
	A	B	C	
1	5	0	0	1
2	1	2	2	0,2
3	0	4	1	0,6
4	0	0	5	1
5	1	4	0	0,6
6	2	3	0	0,4

De posse do resultado de todas as proporções para as mesmas categorias dos 5 médicos, calcula-se  $P_o$  de acordo com a equação (2).

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i = \frac{1}{6} (1 + 0,2 + 0,6 + 1 + 0,6 + 0,4) = 0,6333$$

Agora é necessário calcular as proporções das classificações feitas ao acaso ( $P_e$ ). Para isso, é necessário se calcular antes a equação (6) descrita na tese. Para  $j=1$ , a primeira categoria classificada, temos que:

$$q_i = \frac{1}{5 * 6} \sum_{j=1}^6 x_{1j} = \frac{1}{5 * 6} (5 + 1 + 0 + 0 + 1 + 2) = 0,3$$

Realizando o mesmo procedimento para as outras 3 categorias, teremos a mesma Tabela 3 atualizada para:

Tabela 4. Exemplo do cálculo do coeficiente Kappa para múltiplos observadores com  $q_j$  atualizado.

Exames	Categorias			$P_i$
	A	B	C	
1	5	0	0	1
2	1	2	2	0,2
3	0	4	1	0,6
4	0	0	5	1
5	1	4	0	0,6
6	2	3	0	0,4
Total	9	13	8	
$q_j$	0,30	0,43	0,26	

$P_e$  pode ser calculado então por:

$$P_e = \sum_{j=1}^t q_j^2 = 0,3 * 0,3 + 0,43 * 0,43 + 0,26 * 0,26 = 0,34$$

Por fim, utilizando a fórmula geral de Kappa da equação (1) deste apêndice, teremos:

$$k = \frac{P_o - P_e}{1 - P_e} = \frac{0,63 - 0,34}{1 - 0,34} = 0,44$$

Vale ressaltar que o cálculo do Kappa considera, geralmente, um intervalo de confiança de 95%.