CÉSAR ARMANDO BELTRÁN CASTAÑÓN

Análise e reconhecimento digital de formas
biológicas para o diagnóstico automático de
parasitas do gênero *Eimeria*.

Tese apresentada ao Programa
Interunidades em Bioinformática da
Universidade de São Paulo para a
obtenção do grau de Doutor em
Bioinformática

São Paulo
2006

Candidato(a): César Armando Beltrán Castañón.

Título da Tese: Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero *Eimeria*.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada em/...../, considerou o(a) candidato(a):

() Aprovado(a)

() Reprovado(a)

1) Examinador(a) Assinatura:
Nome:
Instituição:

2) Examinador(a) Assinatura:
Nome:
Instituição:

3) Examinador(a) Assinatura:
Nome:
Instituição:

4) Examinador(a) Assinatura:
Nome:
Instituição:

5) Presidente Assinatura:
Nome:
Instituição:

Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero *Eimeria*

César Armando Beltrán Castañón

TESE APRESENTADA
AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU DE DOUTOR
EM
BIOINFORMÁTICA

Área de Concentração: **Bioinformática**
Orientador: **Prof. Dr. Arthur Gruber**
Co-orientador: **Prof. Dr. Luciano da Fontoura Costa**

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES
- São Paulo, dezembro de 2006 -

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Castañón, César Armando Beltrán

Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero *Eimeria*.

César Armando Beltrán Castañón. -- São Paulo, 2006.

Orientador: Arthur Gruber.

Co-orientador: Luciano da Fontoura Costa.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Área de concentração: Bioinformática. Linha de pesquisa: Processamento digital de imagens.

Versão do título para o inglês: Biological shape analysis and digital recognition for the automatic diagnosis of parasites of the genus *Eimeria*.

Descritores: 1. *Eimeria* 2. Análise morfológica 3. Extração de características 4. Classificação de padrões 5. Processamento de imagens 6. Diagnóstico remoto I. Gruber, Arthur II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Bioinformática. III. Título.

Aos meus pais, Esther e Pedro,
neles encontro meu porto seguro.

Aos meus irmãos,
Norman e Anibal (*in memoriam*),
por serem meus melhores amigos.

Às minhas adoradas, Alejandra e Ledy,
por serem a razão de minha vida;

Agradecimentos

Ao Prof. Dr. Arthur Gruber, por sua grande ajuda na orientação. Com ele aprendi a ter uma atitude rigorosa na pesquisa, assim como levar a ciência como uma religião. Aprecio muito a dedicação e a exigência que ele sempre colocou no trabalho.

Ao Prof. Dr. Luciano da Fontoura Costa, por seu apoio na co-orientação do trabalho, e por ter me fornecido não apenas o conhecimento técnico, mas também por me incutir sempre a importância de publicar.

Ao Prof. Dr. Carlos A. de Bragança Pereira, a quem considero como um segundo co-orientador do trabalho, por ter me obrigado a duplicar esforços para assimilar um novo conhecimento e enriquecer os resultados do trabalho em tão pouco tempo. Com ele muito aprendi, e sempre o terei com admiração e como exemplo.

À Prof^a Dr^a Urara Kawazoe (Unicamp, Brasil), ao Dr. Martin W. Shirley (Institute for Animal Health, Reino Unido), ao Dr. Juan Solis (Laboratório Biovet S/A, Brasil), ao Dr. Harry D. Danforth (USDA, EUA), ao Dr. Petr Bedrník (BIOPHARM, República Tcheca) e ao Dr. Michal Pakandl (Academy of Sciences, República Tcheca), pelas amostras de diferentes cepas e espécies de *Eimeria* spp.

A Sandra Fernandez e Jane Silveira Fraga, pela propagação e processamento dos parasitas, e por sempre terem mostrado disponibilidade e apoio ao trabalho.

Aos meus amigos Bruno Travençolo, Silvia e Luis, por terem me auxiliado na correção do texto da minha tese.

Aos meus colegas que foram e/ou são membros do Grupo de Visão Cibernética do Instituto de Física da USP - São Carlos: Bruno, Silvia, Thomas, Edson, Luis, Leandro Carrijo, Matheus, Erbe, Linder, Ricardo, Dani, Renata, Michelle e Marquinho.

À Profa. Dra. Alda Maria B.N. Madeira e aos meus colegas do Laboratório de Biologia Molecular de Coccídias do Instituto de Ciências Biomédicas da USP-São Paulo: Jane, Sandra, Jeniffer, Ales-

sandra, Ursula, Milene, Ricardo, André e Leonardo, e as técnicas Luciana Terumi Nagao e Livia Rodrigues, pela boa convivência e ajuda recebida.

Ao meu caro amigo e colega Alexandre Proietti Viotti, pela sua amizade e por me desconcentrar, toda vez que programava, com seus “pedalas”. Também fico muito grato pela calorosa recepção que tive de sua família, quando estive em Baependi, Minas Gerais.

À Profa. Dra. Agma Juci Traina, por ter me aceito como seu orientando de mestrado, e assim dar-me a oportunidade de vir ao Brasil. Também agradeço sua permanente preocupação pelo desenvolvimento do meu doutorado. Admiro muito a senhora.

Aos meus pais, “mami Esther” e “papi Pedro” que são os que mais admiro, meus queridos irmãos Anibal (*in memoriam*) e Norman, que me estimularam para continuar com o doutorado. Obrigado por estarem sempre me auxiliando.

À minha esposa Ledy e nossa querida filhinha Alejandra, por serem o estímulo e razão da minha vida.

A minha sogra, “mamá Trina”, minhas cunhadas e cunhados, sobrinhos e sobrinhas, que me apoiaram e proporcionaram a tranquilidade necessária, cuidando de minha filha e minha esposa, durante minha permanência no Brasil.

Ao meu tio, Edwin, às minhas tias, Lucy e Ruth, e aos meus primos Medalith, Luis Alberto, Junior, Nandy e Fallon, pela motivação que sempre recebi deles.

Aos companheiros da república, Waldo e Bruno Feres, e amigos de São Carlos, Leonardo Correa, Douglas, Renato, Vinicius, André Balán, Marcelo e Clever.

Aos colegas e amigos da Sociedade Peruana de Computação, Percy, Alex, Patricia, Waldo, Guillermo, Carlos, Yorch, Eduardo Llapa, Eduardo Tejada, Marco, Juan, Jesus, Christian e, especialmente, ao Ernesto e sua esposa Govy, pelos conselhos e apoio durante o período de estudos no Brasil.

A todos os professores e colegas do Programa de Pós-Graduação Interunidades em Bioinformática da Universidade de São Paulo.

À Patricia, Leka, Pinho, Rosemary, Marilucia e Feijão, pela atenção dispensada.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro.

Resumo

Castañón, C.A.B. **Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero *Eimeria*** [tese]. (Biological shape analysis and digital recognition for the automatic diagnosis of parasites of the genus *Eimeria*). São Paulo: Programa Interunidades em Bioinformática da Universidade de São Paulo; 2006.

O gênero *Eimeria* compreende um grupo de protozoários da classe Coccidia que infecta uma grande variedade de hospedeiros. Um total de sete espécies distintas *Eimeria* podem infectar a galinha doméstica causando enterites com graves prejuízos econômicos. A identificação das espécies pode ser feita através da análise microscópica das diferentes características morfológicas dos oocistos, um dos estágios de desenvolvimento do parasita. Alternativamente, ensaios moleculares baseados na amplificação de alvos específicos de DNA também podem ser utilizados. Em ambos os casos, requer-se um laboratório especializado e, principalmente, pessoal altamente treinado. Neste trabalho é relatada uma abordagem computacional para a extração automática de características para a representação da forma das distintas espécies de *Eimeria*. Foram utilizadas imagens digitais do protozoário nas quais aplicou-se técnicas de processamento de imagens e visão computacional para sua representação morfológica, formando três grupos de características: medidas geométricas, caracterização da curvatura, e quantificação da estrutura interna. A morfologia dos protozoários foi representada por um vetor de características constituído por 14 dimensões, o qual constituiu o padrão de entrada para o processo de classificação. Para o reconhecimento dos padrões, foram usados dois classificadores Bayesianos, utilizando-se como funções de verossimilhança a Gaussiana e a de Dirichlet, respectivamente. O primeiro classificador apresentou as melhores taxas de acerto, enquanto o segundo demonstrou melhor desempenho segundo a análise por curvas ROC. Como prova de princípio de que o sistema poderia ser utilizado por usuários leigos para o diagnóstico à distância de parasitas, foi implementado o COCCIMORPH, um sistema de diagnóstico de *Eimeria* em tempo real. O sistema permite o envio de imagens via *web*, assim como o seu pré-processamento e classificação remotos, obtendo-se o resultado do diagnóstico em tempo real. Essa abordagem totalmente integrada e implementada é inédita para o diagnóstico de parasitas. Entre suas vantagens

principais está o fato de que o diagnóstico pode ser obtido sem a necessidade do transporte físico de amostras biológicas para um laboratório de referência, evitando assim riscos de contaminação do ambiente. Para o treinamento do sistema, foram obtidas centenas de micrografias de cada uma das sete espécies de *Eimeria* que infectam a galinha doméstica. Essas imagens também foram usadas para a construção de um banco de acesso público de imagens (*The Eimeria Image Database*). Além disso, a metodologia de diagnóstico foi também aplicada e testada com onze espécies *Eimeria* de coelho doméstico. Com isso, foram gerados dados inéditos de morfometria, micrografias adicionais para o banco de imagens, e um sistema de classificação para esse conjunto adicional de parasitas. Finalmente, foram determinadas as distâncias entre as diferentes espécies de *Eimeria*, calculadas a partir dos dados morfométricos. As árvores de distância revelaram uma topologia muito similar com árvores obtidas a partir da inferência filogenética usando-se marcadores moleculares como o gene 18S de rRNA ou genomas mitocondriais.

Palavras-chave: *Eimeria*, análise de formas, extração de características, reconhecimento de padrões, processamento de imagens, diagnóstico remoto.

Abstract

Castañón, C.A.B. **Biological shape analysis and digital recognition for the automatic diagnosis of parasites of the genus *Eimeria*** [thesis] (Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero *Eimeria*). São Paulo: Programa Interunidades em Bioinformática da Universidade de São Paulo; 2006.

The *Eimeria* genus comprises a group of protozoan parasites that infect a wide range of hosts. A total of seven different *Eimeria* species infect the domestic fowl, causing enteritis with severe economical losses. Species identification can be performed through microscopic analysis of the distinct morphological characteristics of the oocysts, a developmental stage of the parasite. Alternatively, molecular assays based on the amplification of specific DNA targets can also be used. In both cases, a well equipped laboratory and, especially, highly qualified personnel are required. In this work, we report a computational approach for the automatic feature extraction for shape representation of the different *Eimeria* species. Digital images of the parasites were used in order to apply image processing and computational vision techniques for shape characterization. Three groups of morphological features were constituted: geometric measures, curvature characterization, and internal structure quantification. The protozoan morphology was represented by a 14-dimension feature vector, which was used as the input pattern for the classification process. Two Bayesian classifiers were used for pattern recognition, using as a likelihood function the normal and the Dirichlet, respectively. The former classifier presented the best correct classification rates, whereas the latter showed a better performance in ROC curve analyses. As a proof of principle that this system could be utilized by end-users for a long-distance parasite diagnosis, we implemented COCCIMORPH, an integrated system for the real-time diagnosis of *Eimeria* spp. The system presents an interface for image uploading. Image preprocessing and diagnosis are performed remotely and the results displayed in real-time. This fully integrated and implemented system constitutes a novel approach for parasite diagnosis. Among the several advantages of the system, it is noteworthy that no biological sample transportation is required between the farm and the reference laboratory, thus avoiding potential environment contamination risks. To train the system, we used hundreds of micrographs

of each one of the seven *Eimeria* species of domestic fowl. These images were used to compose a public image repository (*The Eimeria Image Database*). In addition, our diagnosis methodology was extended to the eleven *Eimeria* species that infect the domestic rabbit. With this integrated approach, a totally novel set of images and morphometric data of rabbit *Eimeria* were incorporated to the image database and, also to the remote diagnosis system. Finally, distance trees of the distinct *Eimeria* species of domestic fowl were computed from the morphometric data. The trees revealed a very similar topology with trees obtained with molecular phylogenetic markers such as the 18S rRNA gene and mitochondrial genomes.

Keywords: *Eimeria*, shape analysis, feature extraction, pattern recognition, image processing, remote diagnosis.

Sumário

Lista de Abreviaturas	xv
Lista de Símbolos	xvii
Lista de Figuras	xxi
Lista de Tabelas	xxv
1 Introdução	1
1.1 <i>Eimeria</i> spp. e coccidiose aviária	1
1.2 Diagnóstico de coccidiose	3
1.3 Diagnóstico computacional de formas biológicas	4
1.4 Aplicação dos dados morfológicos na filogenia	9
1.5 Objetivos	10
1.5.1 Objetivo geral	10
1.5.2 Objetivos específicos	10
1.6 Contribuições	11
1.7 Organização da tese	12
2 Aquisição e pré-processamento de imagens	15
2.1 Introdução	15
2.2 Amostras de parasitas	15

2.3	Aquisição de imagens	18
2.3.1	Captura de micrografias	18
2.3.2	Isolamento de oocistos	19
2.4	Pré-processamento de imagens	21
2.4.1	Transformação em tons de cinza	22
2.4.2	Equalização da imagem	23
2.4.3	Segmentação por limiarização	24
2.4.4	Detecção do contorno	25
3	Representação de formas	27
3.1	Introdução	27
3.2	Medidas geométricas	29
3.2.1	Área	29
3.2.2	Diâmetros	30
3.2.3	Simetria	32
3.3	Análise da forma por curvatura	34
3.3.1	Curvatura baseada na transformada multiescala de Fourier	35
3.4	Caracterização da estrutura interna	39
3.4.1	Características para extração de textura	39
3.4.2	Matrizes de co-ocorrência	39
3.5	Espaço de características	41
4	Classificação e mineração	43
4.1	Introdução	43
4.2	Classificação Bayesiana	44
4.2.1	Classificação por densidade normal	45
4.2.2	Classificação por métodos não paramétricos	46
4.3	Processo de classificação	48

4.4	Generalização do classificador	50
4.4.1	O problema da dimensionalidade	51
4.4.2	Seleção de características	54
4.4.3	Função-critério	57
4.4.4	Tamanho mínimo do conjunto de treinamento	57
4.4.5	Avaliação do desempenho do classificador	58
5	Análise de distância a partir de dados morfológicos	61
5.1	Introdução	61
5.2	Inferência filogenética a partir de dados da forma	62
5.3	Discretização	64
5.4	Análise de distância usando dados morfológicos	64
6	Resultados	67
6.1	Introdução	67
6.2	Identificação de espécies de <i>Eimeria</i> de galinha	68
6.2.1	Conjunto de dados	68
6.2.2	Classificadores	68
6.2.3	Desempenho do conjunto de treinamento	69
6.2.4	Correlação de características	71
6.2.5	Seleção de características	71
6.2.6	Discriminação de espécies	74
6.2.7	Análise comparativa do desempenho dos classificadores	79
6.3	Identificação de espécies de <i>Eimeria</i> de coelho	83
6.3.1	Tamanho do conjunto de treino e seleção de características	84
6.3.2	Seleção de características	84
6.3.3	Classificação de espécies	88
6.3.4	Desempenho dos classificadores	88

6.4	Sistema de diagnóstico em tempo real	92
6.4.1	Projeto do sistema de diagnóstico em tempo real	92
6.4.2	Interface <i>web</i>	93
6.4.3	Banco de imagens	95
6.4.4	Morfometria dos oocistos	96
6.5	Análise de distância entre espécies baseada em dados morfológicos	98
6.5.1	Matriz de caracteres	98
6.5.2	Árvores de distância e inferência filogenética	98
7	Discussão e conclusões	101
7.1	Discussão	101
7.1.1	Resumo do trabalho	101
7.1.2	Comparação do sistema COCCIMORPH com outros trabalhos	102
7.1.3	Aquisição de imagens	104
7.1.4	Das características extraídas	105
7.1.5	Da classificação	106
7.1.6	Um novo conceito em diagnóstico de parasitas	109
7.1.7	Um repositório de imagens e de dados morfométricos de parasitas	111
7.1.8	Congruência entre distância morfológica e dados de filogenia molecular	111
7.1.9	Perspectivas futuras	114
7.2	Conclusões	117
	Referências Bibliográficas	119

Lista de Abreviaturas

ANN	Rede neural artificial (<i>Artificial Neural Network</i>).
CBIR	Recuperação de imagens baseado no conteúdo (<i>Content Based Image Retrieval</i>)
CCD	Dispositivo de carga acoplada (<i>Charge-Coupled Device</i>)
CMY	Sistema de cores subtrativas (<i>Cyan, Yellow, Magenta</i>)
CON	Contraste
DNA	Ácido desoxirribonucléico (<i>Desoxyribonucleic acid</i>).
ENT	Entropia
FBST	Teste de significancia genuinamente Bayesiano (<i>Full Bayesian Significance Test</i>)
GLCM	Matriz de co-ocorrência de níveis de cinza (<i>Grey Level Co-occurrence Matrix</i>)
HSI	Espaço de cores (<i>Hue, Saturation, Intensity</i>)
ITS1	<i>Internal Transcribed Spacer 1</i>
JPEG	<i>Joint Photographic Experts Group</i>
MDI	Momento da diferença inversa
mRNA	RNA mensageiro (<i>Messenger RNA</i>).
NP-hard	<i>Non-deterministic Polynomial-time hard</i>
NTSC	Sistema analógico de televisão (<i>National Television System Committee</i>)
PCR	Reacção em cadeia da polimerase (<i>Polymerase Chain Reaction</i>)
RGB	Sistema de cores aditivas (<i>Red, Green, Blue</i>)

SBS	Busca seqüencial para trás (<i>Sequential Backward Selection</i>)
SCAR	Região amplificada caracterizada por sequenciamento (<i>Sequence Characterized Amplified Region</i>)
SFS	Busca seqüencial para frente (<i>Sequential Forward Selection</i>)
SMA	Segundo momento angular
SVM	Maquina de suporte vetorial (<i>Support vector machine</i>).
RAPD	Polimorfismo de DNA amplificado ao acaso (<i>Random Amplified Polimorphic DNA</i>)
RNA	Ácido ribonucleico (<i>Ribonucleic acid</i>).
rRNA	RNA ribossomal (<i>Ribosomal RNA</i>)
ROC	Características operativas do receptor (<i>Receiver Operating Characteristics</i>)

Lista de Símbolos

m, n, i	Variáveis
a	Escala
j	Número imaginário, $\sqrt{-1}$
π	Valor pi = 3,141592
$exp(x)$	$x^e = x^{2,71828}$
s	Sinal (discreto ou contínuo)
t	Tempo ou posição
f	Frequência
I_k	Ponto k -ésimo da imagem I
$\mathbf{x}$	Vetor que representa a coordenada (x, y)
$\mathcal{I}(x, y)$	Valor em tons de cinza do <i>pixel</i> (x, y)
$\mathcal{B}(x, y)$	Valor binário do <i>pixel</i> (x, y)
$f_k(\mathbf{x})$	Nível das componentes R , G e B no ponto k
$R_{\mathbf{x}}, G_{\mathbf{x}}, B_{\mathbf{x}}$	Nível das componentes R , G e B no ponto k
L	Valor de limiarização
$x(t)$ e $y(t)$	Coordenadas paramétricas de um ponto t no contorno do objeto
F, F'	Espaços de pontos conectados
$O(N^2)$	Complexidade quadrática
$T[.]$	Função de transformação
μm	micrômetro
μl	microlitro
$ \cdot $	Valor absoluto de um número

Σ	Somatória
Π	Produto
$\cup$	União
E	Esperança
μ	Média
σ	Desvio padrão
Σ	Matriz de covariância
$c(t)$	Representação paramétrica do contorno com componentes $x(t)$ e $y(t)$
$k(t)$	Curvatura
$\dot{x}(t), \dot{y}(t)$	Primeira derivada de $x(t)$ e $y(t)$
$\ddot{x}(t), \ddot{y}(t)$	Segunda derivada de $x(t)$ e $y(t)$
$X(f), Y(f)$	Transformada de Fourier de $x(t)$ e $y(t)$
$\dot{X}(f), \dot{Y}(f)$	Transformada de Fourier de $\dot{x}(t), \dot{y}(t)$
$\ddot{X}(f), \ddot{Y}(f)$	Transformada de Fourier de $\ddot{x}(t), \ddot{y}(t)$
$q(t) \leftrightarrow Q(f)$	Par genérico de Fourier
$\mathfrak{F}^{-1}$	Inversa da transformada de Fourier
$G(f, \sigma)$	Gaussiana de desvio padrão σ definida no espaço das frequências f
$g(t, \sigma)$	Gaussiana de desvio padrão σ definida no espaço do tempo t
$N(\mu_i, \Sigma_i)$	Distribuição normal definida por μ_i e Σ_i
$\Gamma()$	Função Gamma
δ	Função delta Dirac
$B()$	Função Beta
$X_s(f), Y_s(f)$	Espectros suavizados de $X(f)$ e $Y(f)$
C_{ij}	Probabilidade condicional conjunta dos níveis de cinza i, j
P_{ij}	Frequência de ocorrência entre dois níveis de cinza, i e j
dx, dy	Deslocamento em x e y
d	Valor de deslocamento
g	Número de níveis de cinza quantificados $g = 256$

$\mathbf{x}$	Vetor de características
c	Número de classes
ω_i	Classe i
$P(\omega_j)$	Probabilidade a priori da classe ω_j
$P(\omega_j \mathbf{x})$	Probabilidade a posteriori do vetor $\mathbf{x}$ em relação a ω_j
$p(\mathbf{x} \omega_j)$	Função de verossimilhança de ω_j em relação a $\mathbf{x}$
$p(x)$	Fator de evidência
R_i, R_j	Regiões definidas por funções discriminantes
$g_i(\mathbf{x})$	Função discriminante i em relação a $\mathbf{x}$
d	Dimensão do vetor $\mathbf{x}$ ou número de características
$\mathcal{Y}, \mathcal{X}$	Conjuntos de características
$J(.)$	Função-critério
E	Taxa ou probabilidade de erro
$\mathbf{Y}$	Positivos
$\mathbf{N}$	Negativo
W	Conjunto de todas as classes
P_i	Positivos da classe ω_i
N_i	Negativos da classe ω_i

Lista de Figuras

1.1	Ciclo de vida da <i>Eimeria</i> spp. Arte gráfica: Helton Barreiro.	2
1.2	Fotomicrografias de oocistos das sete espécies de <i>Eimeria</i> da galinha doméstica. (a) <i>E. maxima</i> , (b) <i>E. brunetti</i> , (c) <i>E. tenella</i> , (d) <i>E. necatrix</i> , (e) <i>E. praecox</i> , (f) <i>E. acervulina</i> , e (g) <i>E. mitis</i>	3
1.3	Fluxograma do processo de análise e reconhecimento de imagens de oocistos.	8
2.1	Seqüência da etapa de aquisição e pré-processamento de imagens.	16
2.2	Efeito da resolução de captura na qualidade da imagem do oocisto. Um oocisto foi fotografado em 4 diferentes resoluções sob o mesmo aumento. Uma lâmina microscópica contendo uma escala calibrada (Nikon Inc., USA), foi utilizada para determinar a escala (<i>pixels/μm</i>) de cada resolução. Uma área ampliada da imagem (a) contendo a parede do oocisto e o grânulo polar foi demarcada (retângulo tracejado) e ampliada (b-e) para demonstrar melhor o efeito de serrilhado sob baixas resoluções. Resoluções de captura: (a,b) 2272×1704 (11,1 <i>pixels/μm</i> , 3,9 <i>megapixels</i>); (c) 1600×1200 (8,0 <i>pixels/μm</i> , 1,9 <i>megapixels</i>); (d) 1280×960 (6,4 <i>pixels/μm</i> , 1,2 <i>megapixels</i>); e (e) 1024×768 (5,1 <i>pixels/μm</i> , 0,8 <i>megapixels</i>). As barras de escala correspondendo a $1\mu m$ estão apresentadas no canto superior direito das imagens ampliadas.	19
2.3	Variações na morfologia dos oocistos apresentadas durante a aquisição de imagens. Uma imagem adequada deve estar bem focada e apresentar (a) uma forma normal do oocisto. Algumas variações comuns que impedem o pré-processamento e/ou uma discriminação acurada incluem (b) oocistos amassados, (c) oocistos encolhidos, (d) contraste irregular ou artefatos na parede do oocisto, (e) oocistos mal focados, e (f) oocistos mal posicionados.	20

2.4	Equalização de imagem por “casamento” de histograma. A imagem original (a) foi capturada em condições de baixa luminosidade. Utilizando-se um histograma de uma imagem padrão (b), procedeu-se a uma transformação do histograma da imagem original de forma a “casar” com o histograma da imagem padrão, resultando assim numa imagem equalizada.	24
2.5	Diferentes etapas da rotina de pré-processamento de uma image de oocisto: (a) imagem original colorida, (b) imagem em níveis de cinza, (c) imagem segmentada (binarizada) e (d) detecção do contorno.	25
2.6	Processo de detecção de contorno através do algoritmo de seguimento de contorno (<i>contour following</i>). Os códigos de direção (<i>chain-code directions</i>) estão indicados. .	26
3.1	Cálculo dos diâmetros do objeto baseado nas componentes principais. (a) Objeto em posição original e seus componentes principais, (b) translação do objeto à origem baseado no centro de massa, e (c) rotação do objeto através do alinhamento das componentes principais com os eixos cartesianos, e posterior cálculo dos diâmetros. .	32
3.2	Cálculo da simetria baseado nas componentes principais. Depois que as componentes tenham sido alinhadas com os eixos cartesianos (a), o objeto é rotacionado em função do eixo maior (b) e do eixo menor (c). Os cálculos são feitos sobre a imagem binária (d), a partir da qual são produzidas outras imagens de simetria no eixo maior (e) e no eixo menor (f), onde a região branca representa a porção não simétrica do objeto.	33
3.3	Contorno paramétrico de um oocisto (a) e seu correspondente curvograma usando $\sigma = 10$ (b) e $\sigma = 50$ (c), ou sucessivas variações do desvio padrão na função Gaussiana, mostrada em um curvograma 3D (d).	37
3.4	Contornos paramétricos de <i>E. acervulina</i> (a) e <i>E. maxima</i> (c) e seus respectivos curvogramas (b) e (d) usando $\sigma = 10$	38
3.5	Micrografias (imagens superiores) de oocistos de <i>E. mitis</i> (a), <i>E. brunetti</i> (b) <i>E. maxima</i> (c) e suas respectivas matrizes de co-ocorrência (imagens inferiores) geradas com $d = 2$ e orientação de 90°	40
4.1	Fluxo de trabalho do processo de classificação utilizado neste trabalho.	48
4.2	Fluxograma do processo de generalização do classificador.	52
4.3	Efeito da dimensionalidade na taxa de acerto de um classificador.	53

4.4	Esquema de classificação dos métodos de seleção de características. Adaptada de Jain e Zongker (1997).	55
4.5	Exemplos de curvas ROC.	59
5.1	Geração da matriz de distância a partir de dados morfométricos. Nesse exemplo, a discretização foi feita com intervalos de mesmo tamanho (a), gerando-se sete partições (b), aqui denominadas <i>P1</i> a <i>P7</i> . As médias das medidas morfométricas de cada caractere (c), por espécie, foram designadas para as diferentes partições (b). A partir dessa classificação, foi gerada uma matriz de caracteres (d). A comparação de todas as combinações de pares de espécies foi usada para gerar uma matriz triangular de distância (e).	65
6.1	Efeito do tamanho do conjunto de treino na acurácia da classificação. Um total de 2240 imagens foram usadas na avaliação. O tamanho do conjunto de treino é representado pela porcentagem relativa ao total do conjunto de dados. O número absoluto de imagens também é apresentado (entre parênteses). Os resultados estão apresentados para o classificador por similaridade (linha cheia) e por probabilidade (linha tracejada).	70
6.2	Comparação das taxas de acerto dos classificadores por similaridade (<i>Cs</i>) e probabilístico (<i>Cp</i>) frente ao conjunto de 3891 imagens das sete espécies de <i>Eimeria</i> de galinha doméstica. A avaliação foi feita utilizando-se os métodos de “Todos contra Todos” (<i>Cs</i> -1 e <i>Cp</i> -1), “Partição aleatória” de 30% (<i>Cs</i> -2 e <i>Cp</i> -2) e <i>Leave One Out</i> (<i>Cs</i> -3 e <i>Cp</i> -3).	75
6.3	Comparação dos resultados de classificação para alguns elementos de <i>E. praecox</i> e <i>E. necatrix</i> (http://puma.icb.usp.br/coccimorph/classification/).	80
6.4	Avaliação comparativa dos classificadores por similaridade (linha azul) e por probabilidade (linha vermelha) através de curvas ROC. A linha preta indica a diagonal de referência. Espécies: (a) <i>E. acervulina</i> , (b) <i>E. maxima</i> , (c) <i>E. brunetti</i> , (d) <i>E. mitis</i> , (e) <i>E. praecox</i> , (f) <i>E. tenella</i> , e (g) <i>E. necatrix</i>	81
6.5	Micrografia de oocistos das onze espécies de <i>Eimeria</i> de coelho. Espécies: (a) <i>E. exigua</i> , (b) <i>E. perforans</i> , (c) <i>E. piriformis</i> , (d) <i>E. flavescens</i> , (e) <i>E. irresidua</i> , (f) <i>E. stiedai</i> , (g) <i>E. intestinalis</i> , (h) <i>E. media</i> , (i) <i>E. vej dovskyi</i> , (j) <i>E. coecicola</i> e (k) <i>E. magna</i>	83

- 6.6 Efeito do tamanho do conjunto de treino na acurácia da classificação em *Eimeria* de coelho. Um total de 2167 imagens foram usadas na avaliação. O tamanho do conjunto de treino é representado pela porcentagem relativa ao total do conjunto de dados. O número absoluto de imagens também é apresentado (em parênteses). Os resultados estão apresentados para o classificador por similaridade (linha cheia) e probabilístico (linha tracejada). 86
- 6.7 Avaliação comparativa dos classificadores por similaridade (linha azul) e por probabilidade (linha vermelha) através de curvas ROC. A linha preta indica a diagonal de referência. Espécies: (a) *E. coecicola*, (b) *E. exigua*, (c) *E. flavescens*, (d) *E. intestinalis*, (e) *E. irresidua*, (f) *E. magna*, (g) *E. media*, (h) *E. perforans*, (i) *E. piriformis*, (j) *E. stiedai* e (k) *E. vejnovskyi*. 91
- 6.8 Fluxo de trabalho do sistema integrado de diagnóstico em tempo real (COCCIMORPH). 92
- 6.9 Interface do sistema de diagnóstico em tempo real (COCCIMORPH). (a) Tela principal com imagem de um oocisto enviada ao sistema. O contorno, definido por linha branca circundando o oocisto, pode ser visto. Ao se pressionar o botão *Classify*, o sistema processa a imagem e gera o resultado da classificação em uma nova tela (b). . 94
- 6.10 Tela capturada do sítio web do “*Eimeria Image Database*”, apresentando imagens de oocistos isolados da cepa H de *E. praecox*. 96
- 6.11 Cladogramas gerados (a) com dados morfométricos e (b) por inferência filogenética utilizando-se genomas mitocondriais completos com máxima verossimilhança e o modelo de substituição de nucleotídeos GTR+I+G. 99

Lista de Tabelas

2.1	Origem geográfica das cepas de <i>Eimeria</i> e espécies usadas nesta pesquisa, e o respectivo número de imagens utilizadas neste trabalho.	17
2.2	Equivalência da resolução linear em <i>pixels/μm</i> para algumas resoluções de captura das imagens. Os efeitos da resolução de captura podem ser vistos na Figura 2.2. . . .	18
6.1	Características morfológicas extraídas automaticamente das imagens de <i>Eimeria</i> spp. e seus respectivos identificadores.	69
6.2	Média das matrizes de correlação das sete espécies de <i>Eimeria</i> de galinha	72
6.3	Seleção de características usando SFS e o classificador por similaridade para 3891 elementos	73
6.4	Seleção de características usando SFS e o classificador por probabilidade para 3891 elementos	74
6.5	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por similaridade e validação pelo método “Todos contra Todos” (Cs-1). Média da diagonal = 86,26%.	77
6.6	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por similaridade e validação por “partição aleatória” de 30% (Cs-2). Média da diagonal = 84,87%.	77
6.7	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por similaridade e validação pelo método “ <i>leave-one-out</i> ” (Cs-3). Média da diagonal = 85,49%.	77

6.8	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método “Todos contra Todos” (Cp-1). Média da diagonal = 80,78%.	78
6.9	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método de “partição aleatória” de 30% (Cp-2). Média da diagonal = 79,26%.	78
6.10	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método “leave-one-out” (Cp-3). Média da diagonal = 80,10%.	78
6.11	Origem geográfica das cepas de <i>Eimeria</i> de coelho e respectivo número de imagens utilizadas neste trabalho.	84
6.12	Média das matrizes de correlação das onze espécies de <i>Eimeria</i> de coelho	85
6.13	Seleção de características usando SFS e o classificador por similaridade para as onze espécies de <i>Eimeria</i> de coelho.	87
6.14	Seleção de características usando SFS e o classificador por probabilidade para as onze espécies de <i>Eimeria</i> de coelho.	87
6.15	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. de coelho utilizando-se um classificador por similaridade e validação pelo método “leave-one-out”. Média da diagonal = 80,16%.	89
6.16	Matriz de confusão de diferenciação de espécies de <i>Eimeria</i> spp. de coelho utilizando-se o classificador por probabilidade e validação pelo método “leave-one-out”. Média da diagonal = 73,24%.	90
6.17	Dados morfométricos de <i>Eimeria</i> spp. de galinha doméstica. Os valores mínimos (Mín) e máximos (Máx) dos diâmetros maior (D) e menor (d), assim como os respectivos desvios padrão (D.P.) estão apresentados. O <i>shape index</i> (Razão D/d) também está apresentado.	97
6.18	Dados morfométricos de <i>Eimeria</i> spp. de coelho doméstico. Os valores mínimos (Mín) e máximos (Máx) dos diâmetros maior (D) e menor (d), assim como os respectivos desvios padrão (D.P.) estão apresentados. O <i>shape index</i> (Razão D/d) também está apresentado.	97
6.19	Matriz de caracteres de dados morfométricos	98

Capítulo 1

Introdução

1.1 *Eimeria* spp. e coccidiose aviária

O filo Apicomplexa compreende um grande número de protozoários, sendo a maioria intracelular obrigatória. Esses organismos são caracterizados por apresentar o complexo apical, um conjunto de organelas que está envolvido com a adesão e penetração na célula hospedeira (Current et al., 1990). Fazem parte deste filo alguns dos mais conhecidos e amplamente difundidos patógenos humanos, como os organismos dos gêneros *Plasmodium* spp., agente causador da malária, *Toxoplasma* spp., responsável pela toxoplasmose, *Cryptosporidium* spp. e *Cyclospora* spp., que causam enterites severas. Outros gêneros pertencentes a este filo são de extrema importância na medicina veterinária devido aos graves prejuízos causados na produção animal, como é o caso de *Eimeria* spp., *Isospora* spp., *Theileria* spp. e *Babesia* spp.,

A coccidiose da galinha doméstica é uma doença entérica causada por parasitas do gênero *Eimeria*, que colonizam células da mucosa intestinal das aves (Kawazoe, 1993; McDougald e Reid, 1997). O gênero *Eimeria* compreende mais de 900 espécies de protozoários parasitas, os quais podem ser encontrados nos mais diferentes hospedeiros, desde anelídeos até insetos e vertebrados como aves e mamíferos. No entanto, cada espécie de *Eimeria* infecta apenas uma espécie de hospedeiro. A galinha doméstica pode ser infectada por sete espécies distintas de protozoários do gênero *Eimeria* (Long et al., 1976), levando ao desenvolvimento de uma doença entérica que resulta em menor capacidade absorptiva do intestino, menor ganho de peso e, em alguns casos, aumento da mortalidade dos plantéis. Em função disso, a coccidiose tem uma grande importância na avicultura industrial (Allen e Fetterer, 2002) e persiste como um grande fator de preocupação devido às perdas econômicas que provoca na indústria avícola, apesar do número de drogas anticoccidianas e vacinas que podem ser

utilizadas na sua prevenção (Shirley, 1997; Williams, 2002; Chapman et al., 2002). Os custos associados a profilaxia, tratamentos, as perdas devido ao aumento de mortalidade e queda da produção, são estimados em US\$ 800 milhões por ano em todo o mundo (Williams, 1998; Allen e Fetterer, 2002).

Os parasitas do gênero *Eimeria* apresentam um ciclo de vida monoxênico (veja Figura 1.1), iniciado quando um hospedeiro não imune é infectado pela ingestão de um oocisto esporulado (Current et al., 1990; Kawazoe, 1993; McDougald e Reid, 1997). O oocisto, ao ser ingerido, é rompido na moela da ave por trituração mecânica, liberando os esporocistos. No intestino do animal, devido à ação de sais biliares e enzimas proteolíticas como a tripsina, os esporozoítos saem ativamente da casca do esporocisto e penetram nas células epiteliais do intestino. A partir dessa etapa ocorrem vários ciclos intestinais endógenos com estágios assexuais que se multiplicam por fissão múltipla (merogonia ou esquizogonia), seguidos de um ciclo sexual (gamogonia ou gametogonia) que resulta na formação de um oocisto. O oocisto, ao ser liberado no ambiente, sob condições favoráveis de temperatura, oxigênio e umidade, sofre um processo de esporogonia (divisão meiótica seguida de mitose), resultando num oocisto esporulado contendo quatro esporocistos, com dois esporozoítos cada um, o qual representa a forma infectante do parasita.

Figura 1.1: Ciclo de vida da *Eimeria* spp. Arte gráfica: Helton Barreiro.

1.2 Diagnóstico de coccidiose

Um total de sete espécies distintas de *Eimeria* são consideradas patogênicas para a galinha doméstica: *E. acervulina*, *E. brunetti*, *E. maxima*, *E. mitis*, *E. necatrix*, *E. praecox* e *E. tenella* (Long et al., 1976). Uma vez que diferentes espécies podem variar na patogênese, resistência a drogas e outros parâmetros biológicos, uma adequada discriminação é importante para estudos de epizootiologia e biologia populacional. Outra importante aplicação do diagnóstico de espécies está relacionada ao controle de pureza das amostras caracterizadas (cepas) usadas na preparação de vacinas, detectando-se assim, algum tipo de contaminação.

A identificação de espécies tem sido feita usando critérios morfológicos e patológicos, incluindo o tamanho e forma dos oocistos, período de pré-patência, tempo de esporulação, sítio intestinal e características das lesões (Long e Joyner, 1984; Conway e McKenzie, 1991).

Figura 1.2: Fotomicrografias de oocistos das sete espécies de *Eimeria* da galinha doméstica. (a) *E. maxima*, (b) *E. brunetti*, (c) *E. tenella*, (d) *E. necatrix*, (e) *E. praecox*, (f) *E. acervulina*, e (g) *E. mitis*.

A morfologia do oocisto é uma outra característica que permite discriminar as distintas espécies do parasita. A Figura 1.2 apresenta micrografias de oocistos das sete espécies de *Eimeria* de galinha doméstica. Cada espécie apresenta características morfológicas particulares, sendo algumas dessas características comuns em duas ou mais espécies. Por exemplo, podemos verificar que o tamanho de *E. acervulina* é similar ao de *E. mitis*. A estrutura interna de *E. praecox* mostra-se diferenciada das demais espécies, assim como o tamanho de *E. maxima* é maior do que as demais. Por outro lado, *E.*

praecox e *E. mitis* compartilham uma forma circular, *E. maxima* e *E. brunetti* uma forma ovóide, e *E. acervulina* e *E. necatrix* uma forma elíptica.

A identificação de espécies por inspeção direta, seja das lesões, seja dos oocistos, apresenta algumas desvantagens: (1) número pequeno de características distintas, (2) os critérios usados podem ser subjetivos, (3) sobreposição de parâmetros entre distintas espécies e (4) o transporte de amostras biológicas das granjas comerciais até o laboratório de diagnóstico.

Shirley (1975) foi o primeiro a usar uma abordagem de biologia molecular para diferenciar espécies com base nos padrões eletroforéticos de isoenzimas. No início dos anos 90, Welsh e McClelland (1990) e Williams et al. (1996) desenvolveram a técnica de RAPD (polimorfismo de DNA amplificado ao acaso), ensaio baseado na amplificação de alvos anônimos pelo uso de primers arbitrários. O resultado do ensaio é a geração de um perfil de múltiplas bandas, também conhecido como *fingerprint*, o qual varia entre indivíduos ou populações. Schnitzler et al. (1998, 1999) desenvolveram um ensaio diagnóstico de PCR, baseado na amplificação da região ribossômica ITS1. O nosso grupo também desenvolveu um conjunto de marcadores moleculares denominados SCARs (Fernandez et al., 2003a), os quais permitem o diagnóstico por uma PCR específica. Foi ainda desenvolvido um teste de PCR multiplex no qual o diagnóstico das sete espécies pode ser obtido por meio de uma reação simples em um único tubo (Fernandez et al., 2003b).

Embora as técnicas diagnósticas moleculares tenham uma alta sensibilidade e especificidade, elas requerem pessoal especializado. Além disso, é necessário o transporte de amostras de fezes das aves da granja até o laboratório de referência, o que pode apresentar riscos sanitários.

Na última década houve um enorme avanço da tecnologia e a correspondente queda nos preços de equipamentos de aquisição de imagens e processamento computacional de grandes quantidades de dados. Com isso, técnicas de visão computacional e reconhecimento de padrões passaram a representar uma alternativa potencial para o diagnóstico morfológico de amostras biológicas.

1.3 Diagnóstico computacional de formas biológicas

Uma das aplicações importantes da análise de imagens consiste na classificação e reconhecimento de objetos de interesse em imagens digitais. Os objetos podem ser caracterizados de distintas formas como, por exemplo, identificando-se a cor, textura, forma, movimento e localização do objeto dentro da imagem. Até o momento não se conseguiu desenvolver um método que resolva o problema de reconhecimento automático de padrões para diferentes domínios de imagens. Em um sistema de reconhecimento automático é indispensável se realizar uma análise prévia do domínio de imagens

a serem reconhecidas, pois mesmo o sistema de percepção humana necessita de um treinamento prévio que permita o reconhecimento de sinais. Algumas aplicações de reconhecimento de padrões em problemas biológicos, especialmente para fins de diagnóstico, foram reportados na literatura. Comaniciu et al. (1999) desenvolveram um sistema de recuperação de imagens por conteúdo, o qual permite discriminar linfomas malignos de linfócitos crônicos de leucemia utilizando descritores de textura e forma. Em um trabalho similar para o diagnóstico de leucemia linfóide, Sabino et al. (2004) utilizaram matrizes de co-ocorrência de tons de cinza para caracterizar a textura dos linfóides. Jalba et al. (2006) propuseram uma abordagem para a identificação automática de diatomáceas, a qual está baseada na análise do contorno através da construção de um espaço de curvatura morfológica para a extração de características. Assim como é importante a extração das características das imagens biológicas, o tipo de classificador utilizado também é fundamental no processo de reconhecimento de padrões. O método mais usado para formas encontradas na natureza é a análise multivariada baseada em distribuição Gaussiana, a qual foi utilizada com sucesso na identificação de tipos de bactérias (Trattner et al., 2004), reconhecimento de culturas celulares (Long et al., 2005), e classificação de imagens de cromossomos (Sampat et al., 2005).

Uma interessante área de aplicação para a implementação de sistemas de identificação baseados em imagens é o diagnóstico de parasitas. Os parasitas são geralmente discriminados e identificados através de análises morfológicas não automatizados (inspeção visual macro- e/ou microscópica), ou através de técnicas de biologia molecular. Considerando-se que a grande maioria dos parasitas apresenta estágios de desenvolvimento com uma morfologia bem definida e homogênea, eles constituem um interessante modelo de estudo para técnicas de reconhecimento de padrões. A *Eimeria* spp. apresenta um estágio denominado oocisto (vide item 1.1), o qual apresenta estrutura arredondada e cuja morfologia varia entre as espécies quanto ao tamanho (área, diâmetros), forma do contorno (elíptico, oval, circular), estrutura interna, espessura da parede, cor, entre outras variações morfológicas (Figura 1.2). Em função disso, esse parasita tem sido um dos mais empregados em estudos de análise e reconhecimento de imagens (Kucera e Reznicky, 1991; Dauschies et al., 1999; Plitt et al., 1999).

Vários estudos têm sido relatados na literatura para a diferenciação de espécies de *Eimeria* (Kucera e Reznicky, 1991; Dauschies et al., 1999; Plitt et al., 1999) e de helmintos (Joachim et al., 1999) através do uso de reconhecimento de imagens digitais. O trabalho de Kucera e Reznicky (1991) foi um dos primeiros a usar imagens digitais para a diferenciação das espécies de *Eimeria* de galinha doméstica, mas ele está restrito ao uso de apenas duas características (diâmetro maior e diâmetro menor do oocisto), as quais foram calculados de forma semi-automática. Conforme discutido no item 1.2, a sobreposição de dados morfométricos limita a diferenciação de todas as espécies. Sommer (1998a,b), trabalhando com *Eimeria* de bovinos, usou uma abordagem mais

complexa, onde o contorno paramétrico foi utilizado como entrada para calcular a amplitude da transformada de Fourier. No entanto, o método de classificação aplicado nesse trabalho (*average linkage clustering*) usa como métrica a distância Euclideana, a qual não leva em consideração a distribuição dos elementos, além de estar voltado ao agrupamento dos elementos, técnica não muito indicada para o desenvolvimento de um processo automático de identificação. Yang et al. (2001) desenvolveram um sistema automático para detectar e classificar ovos de helmintos usando redes neurais artificiais (ANNs). Os autores usaram como características a amplitude da transformada de Fourier calculada a partir do contorno paramétrico do objeto, usado por Sommer (1998a). Os resultados de validação cruzada mostraram alta porcentagem de classificação correta, variando de 86,1 a 90,3%, mas o pequeno número de amostras de imagens utilizadas não permitiu uma estimativa segura do nível de confiança dessa abordagem. Widmer et al. (2002) também descreveram o uso de redes neurais artificiais para a detecção de oocistos de *Cryptosporidium parvum*. Os autores conseguiram diferenciar com sucesso os oocistos dos debris presentes na imagem, mas não chegaram a desenvolver uma diferenciação de espécies. Ross et al. (2006) apresentaram uma abordagem de processamento de imagens para o diagnóstico e discriminação de espécies do gênero *Plasmodium*. Como esses parasitas infectam as hemáceas dos pacientes, o primeiro desafio foi o de discriminar as células infectadas daquelas não infectadas. Essa discriminação foi obtida com certo êxito, porém, a determinação da espécie do parasita não foi eficiente. Uma das limitações para essa discriminação foi a dificuldade de se segmentar adequadamente os parasitas dentro das hemáceas, visto que a morfologia do parasita é fundamental para a determinação da espécie.

Uma das maiores dificuldades na análise de formas está na quantificação morfológica, o que explica o limitado número de características usadas nos distintos trabalhos revisados. Esta limitação, junto com a alta complexidade dos algoritmos, faz com que o desenvolvimento de sistemas de diagnóstico automático em tempo real seja uma tarefa muito desafiadora. Além disso, as características a serem utilizadas são fortemente dependentes da especificidade do domínio de imagens. Nesse sentido, nosso grupo de Visão Cibernética tem desenvolvido técnicas para a análise e classificação de formas (Costa e Cesar Jr., 2000). Assim, Bruno et al. (1998) usaram características multiescala para a representação de células neurais ganglionares do gato, enquanto que Coelho et al. (2002) propuseram outro conjunto de características (diâmetro, excentricidade, dimensão fractal, histogramas de influência, área de influência, área e diâmetro do *convex hull*) para o mesmo problema. Costa et al. (2004) fizeram uso da curvatura digital para diferenciação morfológica de crânios da espécie roedor *Thrichomys apereoides*.

A análise de imagens biológicas é uma das múltiplas aplicações da visão computacional que, junto a muitas outras aplicações, comprovam a importância do sentido da vista para os seres huma-

nos. Embora a ação de ver seja muito natural para os humanos, existem grandes dificuldades na implementação de sistemas computacionais de visão que sejam versáteis e confiáveis. O caráter interdisciplinar da pesquisa em visão ilustra o nível de complexidade necessário para a implementação de uma abordagem, a qual tem que normalmente utilizar técnicas de processamento de imagens, reconhecimento de padrões, processamento de sinais, computação gráfica, análise estatística, biologia, entre outros. Nos últimos anos, modelos matemáticos foram propostos para simular algumas funções da visão, como a detecção de bordas (Marr, 1982), relação entre o conceito biológico de campos receptivos e filtros de Gabor 2D (Daugman, 1988), modelos artificiais de redes neurais (Anderson, 1995), importância da percepção humana de vértices e pontos de alta curvatura em formas (Attneave, 1954). Com o incremento do poder computacional, muitas das técnicas de visão têm sido implementadas para trabalhar com alguma eficiência, mas até o momento não foi possível atingir a performance em tempo real do sistema de visão biológico.

A classificação de padrões tornou-se uma ferramenta central na bioinformática (Liew et al., 2005), facilitando o tratamento de grandes conjuntos de dados (Baldi e Brunak, 1998). O uso da classificação de padrões tem-se diversificado em distintas áreas de aplicação biológica, como análise de *microarray* (Valafar, 2002), caracterização estrutural e funcional de proteínas (Chou e Zhang, 1995; Shen e Chou, 2006), predição de genes (Xu e Uberbacher, 1996) e caracterização de imagens médicas, entre outras.

Um grande número de sistemas existentes estão orientados a trabalhar na recuperação de imagens por conteúdo (CBIR – *Content Based Image Retrieval*), onde imagens são recuperadas por técnicas de similaridade, baseadas numa imagem de consulta (Gudivada e Raghavan, 1995; Smeulders et al., 2000; Veltkamp e Tanase, 2000; Müller et al., 2004). O autor, no seu trabalho de mestrado, desenvolveu um sistema CBIR para a recuperação por conteúdo em um banco de imagens médicas através de *wavelets* (Castanón, 2003). Geralmente, a recuperação em sistemas CBIR não implica na diferenciação de classes dentro de um domínio de imagens. A motivação disso é que as características usadas são globais e utilizam classificadores não supervisionados pelo fato de não trabalhar com domínios específicos de imagens.

Por outro lado, os sistemas de reconhecimento de padrões supervisionados precisam de protótipos que vão ser usados como conjunto de treino, o que implica em se ter um conjunto mínimo de exemplos para cada classe envolvida. A Figura 1.3 apresenta um fluxograma do processo geral de análise e reconhecimento de imagens, sendo, neste exemplo, aquele empregado no presente trabalho. Esse fluxograma segue um modelo usual em sistemas de reconhecimento de padrões, compreendendo basicamente três etapas: (1) pré-processamento, (2) extração de características (transformação) e (3) reconhecimento de padrões (classificação) (Duda et al., 2001; Costa e Cesar Jr., 2000).

Figura 1.3: Fluxograma do processo de análise e reconhecimento de imagens de oocistos.

O desenvolvimento de sistemas de tempo real para o reconhecimento automático de imagens tem avançado de forma lenta em comparação com a evolução dos equipamentos de aquisição de imagens, os quais estão cada vez mais baratos, com maior resolução, e com maior número de características embutidas (Minkel, 2006; Ashley, 2006). Assim, a rápida popularização das câmeras digitais, juntamente com a expansão da internet, torna viável a criação de sistemas de diagnóstico remoto em tempo real.

1.4 Aplicação dos dados morfológicos na filogenia

A morfologia é a abordagem clássica para classificar as distintas espécies de organismos. A clássica teoria de Darwin sobre a origem e evolução das espécies (Darwin, 1859) apoiou-se em observações de mudanças morfológicas para evidenciar a adaptação das espécies ao meio ambiente, constituindo uma nova área, a biologia evolutiva, que estuda a origem e a descendência das espécies, bem como suas mudanças ao longo do tempo, ou seja, sua evolução. A história evolutiva de uma espécie, que descreve as várias espécies das quais ela descende, juntamente com a sua relação com outras espécies vivas, constituem a sua filogenia.

A inferência filogenética tem sido classicamente feita usando-se características morfológicas, especialmente em organismos complexos. Mas, com o desenvolvimento de técnicas de sequenciamento de DNA, os estudos filogenéticos passaram a utilizar seqüências de DNA ou proteínas. Os estudos mais comuns são feitos com base em seqüência de nucleotídeos e aminoácidos de genes altamente conservados, como os de RNA ribossômico (subunidade 18S), genes codificadores de enzimas e proteínas estruturais, como o gene nuclear da β -actina e, ainda, genes exclusivamente mitocondriais como o de citocromo *b*. O uso de marcadores moleculares permitiu a reconstrução filogenética de microorganismos, tarefa que seria praticamente impossível de realizar apenas com base nas características morfológicas. Além disso, o uso de seqüências biológicas para a inferência filogenética está atualmente embasada em modelos de evolução relacionados à substituição de bases no DNA. Por outro lado, o uso de dados morfométricos para inferências de relações evolutivas é palco de grandes controvérsias, com alguns autores propondo que não sejam usados em estudos filogenéticos (Scotland et al., 2003), e outros defendendo o seu uso (Jenner, 2004; Wiens, 2004). A maior crítica à reconstrução filogenética com base morfológica é o número limitado de características morfológicas que podem ser utilizadas para inferir as relações evolutivas entre espécies.

Em espécies de *Eimeria*, foram feitas até o momento reconstruções filogenéticas utilizando-se apenas seqüências 18S ribossômicas (Barta et al., 1997, 1998). Nosso grupo, utilizando seqüências completas dos genomas mitocondriais das sete espécies de *Eimeria* de galinha doméstica, realizou a sua reconstrução filogenética (Romano, 2004) e obteve resultados bastante similares àqueles descritos para o rRNA 18S. Uma alternativa interessante, utilizando-se dados de morfologia, seria o de comparar a distância baseada em morfometria, entre as espécies, com aquela observada em reconstruções filogenéticas baseadas em marcadores moleculares. Assim, o uso de técnicas de processamento de imagens, juntamente com as de reconhecimento de padrões, pode gerar potenciais ferramentas para a solução de vários problemas biológicos não diretamente relacionados, como o diagnóstico de espécies de um lado, e a relação evolutiva entre elas, por outro.

1.5 Objetivos

1.5.1 Objetivo geral

Esta tese tem como objetivo desenvolver uma abordagem para o diagnóstico de espécies de parasitas do gênero *Eimeria* mediante a caracterização morfológica e análise de imagens digitais de oocistos, aplicando técnicas de reconhecimento de padrões e visão computacional.

1.5.2 Objetivos específicos

Dentro do caráter multidisciplinar do trabalho de pesquisa desenvolvido, e no intuito de atingir nosso objetivo geral, foram definidos alguns objetivos específicos:

- Definir um conjunto de características que identifiquem a morfologia das distintas espécies de *Eimeria* de galinha.
- Selecionar o conjunto adequado de características que permitam otimizar o processo de diagnóstico.
- Definir o número mínimo de elementos que devem formar o conjunto de treino para o classificador supervisionado.
- Analisar e comparar o desempenho dos classificadores usados para o diagnóstico e diferenciação das distintas espécies.
- Desenvolver um sistema de auxílio ao diagnóstico da coccidiose aviária (causada por parasitas do gênero *Eimeria*) que permita realizar consultas via interface *web* de forma remota e em tempo real.
- Selecionar e montar uma base de dados pública de imagens de oocistos das distintas espécies de *Eimeria* de galinha e de diferentes amostras caracterizadas (cepas) de cada espécie.
- Comparar os resultados de medidas morfológicas obtidas automaticamente com estudos de morfometria já existentes na literatura.
- Usar as medidas morfológicas em análises de distância e comparar os resultados com os obtidos por métodos de filogenia molecular.
- Propor um protocolo para estender nossa abordagem para o diagnóstico de outros parasitas.

- Testar a abordagem desenvolvida com outras espécies de *Eimeria*, especificamente, com aquelas que infectam o coelho.

1.6 Contribuições

O presente trabalho apresenta uma nova abordagem no diagnóstico de parasitas, usando *Eimeria* spp. como modelo e prova do conceito. Para isso, utiliza-se a análise automática de imagens para a caracterização morfológica dos oocistos, um estágio de desenvolvimento do parasita facilmente encontrado. As características morfológicas podem ser aproveitadas para outras aplicações, incluindo a morfometria e a análise de distância entre espécies, esta última permitindo comparações com inferências filogenéticas com marcadores moleculares. A metodologia proposta pode ser estendida a outros parasitas e/ou microorganismos, ampliando assim o escopo do método proposto. As contribuições do trabalho podem ser divididas genericamente nos seguintes aspectos:

1. **Conjunto de características para a representação morfológica de microorganismos do gênero *Eimeria* spp.:** Este trabalho propõe um conjunto de características para a representação da morfologia de parasitas do gênero *Eimeria*. No total, foram extraídas 14 características das imagens dos oocistos, as quais foram divididas em três grupos: medidas geométricas, caracterização da curvatura e representação da estrutura interna do organismo. Essas características foram automaticamente extraídas a partir das imagens digitais, utilizando-se técnicas de processamento de imagens e de visão computacional.
2. **Classificação por similaridade e por probabilidade:** As características extraídas foram submetidas a uma função de classificação que faz a discriminação das distintas espécies. Neste trabalho foram usados dois classificadores, ambos baseados na abordagem Bayesiana: um usa a função de densidade normal fornecendo resultados interpretados como de similaridade, enquanto o outro usa a função de Dirichlet para calcular resultados por probabilidade. Embora o primeiro classificador (por similaridade) apresente-se como o de melhor taxa de acerto, depois de uma análise comparativa de classificadores através de curvas ROC, o segundo classificador (por probabilidade) mostrou-se com melhor desempenho. Sistemas de diagnóstico dificilmente fornecem uma resposta definitiva e com 100% de acurácia, de forma que interpretação do especialista humano ainda é fundamental.
3. **Implementação de um sistema de diagnóstico em tempo real:** Foi desenvolvido um sistema de diagnóstico (COCCIMORPH) que permite o envio de imagens via *web*, assim como o

pré-processamento e diagnóstico remotos da imagem, com geração do diagnóstico em tempo-real. Essa abordagem é um novo conceito no diagnóstico de parasitas, permitindo o diagnóstico a longa distância, sem a necessidade de se transportar amostras biológicas para um laboratório de referência. A página do sistema, contendo a interface de usuário, bem como a documentação sobre a funcionalidade do sistema pode ser acessada no endereço: <http://puma.icb.usp.br/coccimorph>.

4. **Análise de distância a partir de dados morfológicos:** A reconstrução filogenética usando dados morfológicos tem sido feita geralmente para organismos complexos. Com a popularização das técnicas de seqüenciamento de DNA, esse procedimento geralmente tem sido feito através da comparação de seqüências de DNA ou proteínas. No presente trabalho são mostrados os resultados de análises de distância entre espécies de *Eimeria* e sua comparação com árvores filogenéticas determinadas com marcadores moleculares. A concordância de topologias das árvores, e a própria comparação entre métodos baseados em marcadores moleculares e morfológicos é inédita em microrganismos.
5. **Banco de dados de imagens de *Eimeria*:** O amplo conjunto de micrografias de oocistos foi também utilizado para compor um banco público de imagens, denominado “*The Eimeria Image Database*”. Também estão disponibilizados os distintos conjuntos de características, constituindo-se num repositório de dados que pode ser usado para várias finalidades: treinamento de pessoal para identificação dos parasitas, utilização das imagens para testes de classificadores, seleção de características ou de análises de distância.

1.7 Organização da tese

O Capítulo 2 apresenta a aquisição de imagens a partir de amostras biológicas, e as técnicas utilizadas no pré-processamento das imagens. Esse capítulo começa com a descrição das distintas amostras usadas nos experimentos, a seguir apresenta o processo de aquisição de imagens para, em seguida discutir as técnicas de pré-processamento que compreende a transformação em tons de cinza, equalização e segmentação das imagens e, finalmente, a detecção do contorno do objeto em análise.

O Capítulo 3 introduz as distintas técnicas usadas na caracterização da forma dos oocistos, as quais compreendem técnicas de extração de características que implicam em (a) extração de medidas geométricas, (b) análise da forma por curvatura e (c) análise da estrutura interna dos oocistos através da caracterização da textura. O capítulo finaliza propondo um espaço de características a ser utilizado

nos experimentos.

O Capítulo 4 apresenta os métodos de classificação usados nos experimentos para o diagnóstico de espécies. Para isso, foi usada a metodologia Bayesiana, mas com duas abordagens diferentes que denominamos de (a) classificação por similaridade e (b) classificação por probabilidade. Acompanham esse capítulo, a descrição dos algoritmos de classificação usados nesse trabalho, a técnica de seleção de características e a metodologia usada para comparar o desempenho dos classificadores.

O Capítulo 5 apresenta os métodos usados na análise de distância de espécies de *Eimeria* usando-se os dados morfológicos. Nesse capítulo são abordadas as técnicas de discretização dos dados morfológicos e a geração de dendrogramas.

No Capítulo 6 apresentam-se os resultados obtidos nos distintos experimentos para cada uma das etapas envolvidas no projeto. Inicia-se com a apresentação dos resultados da discriminação de espécies de *Eimeria* de galinha (Seção 6.2) definindo o conjunto adequado de características que maximizam a correta classificação de oocistos. Também foi determinado o número mínimo de imagens de oocistos que pode ser usado para treinar os classificadores e, posteriormente, é apresentada uma análise comparativa dos classificadores usados. Também são apresentados os resultados obtidos para as espécies de *Eimeria* de coelho (Seção 6.3), como comprovação da validade da nossa abordagem. O sistema de diagnóstico em tempo real é introduzido na Seção 6.4. A arquitetura, funcionalidade, uso e adaptação para outros parasitas são explicados. Finalmente, a Seção 6.5 apresenta resultados inéditos de análise de distância, baseada em caracteres morfológicos dos parasitas. Os resultados são comparados com os obtidos através de métodos de filogenia molecular.

O Capítulo 7 apresenta uma discussão sobre alguns possíveis desdobramentos do trabalho desenvolvido que podem ser continuados em pesquisas futuras. Finalmente, são apresentadas as conclusões do trabalho.

A seguir apresenta-se um esquema da estrutura desta tese, considerando as relações entre seus capítulos e apêndices.

Capítulo 2

Aquisição e pré-processamento de imagens

2.1 Introdução

Este capítulo aborda as etapas envolvidas no pré-processamento das imagens microscópicas. A Figura 2.1 apresenta a seqüência de passos envolvidos no pré-processamento, a qual compreende duas fases: aquisição de imagens e pré-processamento digital. A primeira fase é um trabalho desenvolvido no laboratório biológico, o que implica na coleta das fezes da galinha, purificação dos oocistos e aquisição de imagens por microscopia.

Na segunda fase, o objetivo é determinar o objeto de interesse a ser trabalhado na etapa de análise de formas (ver Capítulo 3). As micrografias obtidas são inspecionadas visualmente para se isolar aqueles oocistos que não apresentam artefatos na sua morfologia. As imagens isoladas são transformadas em tons de cinza para então serem equalizadas e estarem aptas para o processo de segmentação dos objetos de interesse (binarização). Na imagem binarizada é aplicado o algoritmo de detecção de contorno que permite obter uma representação paramétrica da borda do objeto.

2.2 Amostras de parasitas

Para esse trabalho foram coletadas amostras de parasitas de cada uma das sete espécies de *Eimeria* que infectam a galinha doméstica. Além disso, sempre que possível, foram usadas diversas cepas de cada espécie, coletadas de diferentes regiões geográficas (Tabela 2.1). As galinhas foram infectadas por via oral com oocistos purificados nas doses recomendadas por Shirley e Harvey (1996). A coleta de oocistos, a purificação e a esporulação foram feitas em nosso laboratório seguindo os

Figura 2.1: Sequência da etapa de aquisição e pré-processamento de imagens.

procedimentos padrão (Long et al., 1976). A correta designação de espécies e a pureza da amostra foram confirmadas mediante o procedimento de PCR desenvolvido pelo nosso grupo Fernandez et al. (2003a,b). Os parasitas foram propagados em galinha de 3 a 4 semanas de idade. Os pintos de um dia de idade foram fornecidos por uma granja comercial (Granja Kunitomo, Mogi das Cruzes, SP, Brasil) e receberam cuidados em um ambiente livre de coccidiose e alimentados com água filtrada e ração especial para um crescimento livre de anticoccidianos e antibióticos (Braswey S.A. - Indústria e Comércio, Campinas, SP, Brasil). Todas as propagações e purificações de oocistos foram feitas por Sandra Fernandez e Jane Silveira Fraga, colegas de pós-graduação do nosso laboratório.

Tabela 2.1: Origem geográfica das cepas de *Eimeria* e espécies usadas nesta pesquisa, e o respectivo número de imagens utilizadas neste trabalho.

Espécie	Origem	Nº de amostras
<i>E. acervulina</i> H	Houghton, Inglaterra	374
<i>E. acervulina</i> 103	São Paulo, Brasil	114
<i>E. acervulina</i> R7	Santa Catarina, Brasil	148
<i>E. brunetti</i> C	São Paulo, Brasil	418
<i>E. maxima</i> H	Houghton, Inglaterra	103
<i>E. maxima</i> L	São Paulo, Brasil	91
<i>E. maxima</i> 50	São Paulo, Brasil	127
<i>E. mitis</i> CR	República Tcheca	335
<i>E. mitis</i> 30	São Paulo, Brasil	199
<i>E. mitis</i> 44	São Paulo, Brasil	223
<i>E. necatrix</i> DF	São Paulo, Brasil	259
<i>E. necatrix</i> 103	São Paulo, Brasil	145
<i>E. praecox</i> H	Houghton, Inglaterra	377
<i>E. praecox</i> 1D1A	São Paulo, Brasil	180
<i>E. praecox</i> D	USA	190
<i>E. tenella</i> H	Houghton, Inglaterra	311
<i>E. tenella</i> CR	República Tcheca	137
<i>E. tenella</i> MC	São Paulo, Brasil	160

As cepas de origem brasileira foram purificadas e propagadas em nosso laboratório, a partir de um único oocisto, isolados de cada amostra coletada em granjas comerciais no Estado de São Paulo. A amostra Houghton (H) foi gentilmente cedida pelo Dr. Martin W. Shirley (Institute for Animal Health, Compton, Reino Unido). O isolado *E. mitis* CR, isolado na República Tcheca foi gentilmente cedido pelo Dr. Petr Bedrník (BIOPHARM - Research Institute of Biopharmacy and Veterinary Drugs, República Tcheca).

2.3 Aquisição de imagens

A fase de aquisição envolve a captura digital de uma imagem (fotografia) para que possa ser subsequentemente processada pelo computador. O tipo de dispositivo e as condições de aquisição têm uma forte influência na aplicação final. Para aplicações biológicas, geralmente é usado um microscópio ao qual é acoplada uma câmera digital para se obter as respectivas micrografias. Uma vez que uma micrografia pode conter vários oocistos, o próximo passo é o recorte de oocistos individuais e sua gravação em arquivos de formato JPEG distintos.

2.3.1 Captura de micrografias

As micrografias foram obtidas com um microscópio óptico (Nikon Eclipse E800) acoplado a uma câmera CCD de 4,0 *megapixels* (Nikon Coolpix 4500). As micrografias foram capturadas com uma objetiva de 40 $\times$, com a câmera na posição de zoom 4.1 $\times$, e gravadas em formato JPEG de 24 bits (opção de qualidade fina), com uma resolução de 2272 $\times$ 1704 *pixels* (total de 3,9 *megapixels*).

A Figura 2.2 apresenta os efeitos da resolução de captura na definição das imagens. Considerando-se a parte ampliada da Figura 2.2(a), uma imagem de alta resolução (Figura 2.2(b)) apresenta uma boa definição do contorno do objeto. À medida que se diminui a resolução da imagem, a borda do objeto começa a apresentar o efeito de serrilhado (Figura 2.2(e)).

Nas condições de captura descritas acima, as imagens capturadas em diferentes resoluções tiveram suas resoluções lineares (em *pixels*/ μm) determinadas. Os valores encontrados estão apresentados na Tabela 2.2.

Tabela 2.2: Equivalência da resolução linear em *pixels*/ μm para algumas resoluções de captura das imagens. Os efeitos da resolução de captura podem ser vistos na Figura 2.2.

Tamanho	Resolução	<i>Pixels</i> / μm
2272 $\times$ 1704	3,9 <i>megapixels</i>	11,1
1600 $\times$ 1200	1,9 <i>megapixels</i>	8,0
1280 $\times$ 960	1,2 <i>megapixels</i>	6,4
1024 $\times$ 768	0,8 <i>megapixels</i>	5,1

É importante destacar que alguns fatores podem interferir na morfologia dos oocistos como, por exemplo, o volume de suspensão de oocistos a ser depositado entre a lâmina e a lamínula microscópicas. Inicialmente, trabalhou-se com 3,0 μl , o que permitia uma melhor definição do contorno e da estrutura interna do oocisto, mas verificou-se que os oocistos sofriam alteração na sua morfologia, sendo pressionados pelo peso da lamínula devido à pouca quantidade de líquido.

Figura 2.2: Efeito da resolução de captura na qualidade da imagem do oocisto. Um oocisto foi fotografado em 4 diferentes resoluções sob o mesmo aumento. Uma lâmina microscópica contendo uma escala calibrada (Nikon Inc., USA), foi utilizada para determinar a escala ($pixels/\mu m$) de cada resolução. Uma área ampliada da imagem (a) contendo a parede do oocisto e o grânulo polar foi demarcada (retângulo tracejado) e ampliada (b-e) para demonstrar melhor o efeito de serrilhado sob baixas resoluções. Resoluções de captura: (a,b) 2272×1704 ($11,1\ pixels/\mu m$, $3,9\ megapixels$); (c) 1600×1200 ($8,0\ pixels/\mu m$, $1,9\ megapixels$); (d) 1280×960 ($6,4\ pixels/\mu m$, $1,2\ megapixels$); e (e) 1024×768 ($5,1\ pixels/\mu m$, $0,8\ megapixels$). As barras de escala correspondendo a $1\mu m$ estão apresentadas no canto superior direito das imagens ampliadas.

Com isso, os oocistos são achatados e têm o seu tamanho artificialmente aumentado, com evidente distorção de sua morfologia. Pode-se visualizar este artefato comparando-se as imagens das Figuras 2.3(a) e (b) de um oocisto de *E. tenella*. Após tentativas empíricas, concluímos que o volume mais adequado deveria ser de $6,0\mu l$.

2.3.2 Isolamento de oocistos

O processo de recorte (*cropping*) dos oocistos das micrografias foi realizado manualmente através do uso de softwares para edição de imagens. Qualquer programa pode ser usado, como por exemplo o Adobe Photoshop[®], Gimp ou Corel PHOTO-PAINT[®]. Os objetos de interesse, nesse caso oocistos individuais, foram selecionados e recortados da micrografia, e gravados de forma separada em arquivos JPEG de 24 *bits*.

É importante destacar que numa micrografia de um campo microscópico, nem todos os oocistos são adequados para o sistema de reconhecimento. A Figura 2.3 apresenta exemplos de oocistos

isolados de *E. tenella* que, por diferentes razões não são aptos para um processo automático de reconhecimento, com exceção do painel (a). De fato, os oocistos podem apresentar morfologias distorcidas em função de vários artefatos e/ou variáveis não controladas no momento da aquisição.

Figura 2.3: Variações na morfologia dos oocistos apresentadas durante a aquisição de imagens. Uma imagem adequada deve estar bem focada e apresentar (a) uma forma normal do oocisto. Algumas variações comuns que impedem o pré-processamento e/ou uma discriminação acurada incluem (b) oocistos amassados, (c) oocistos encolhidos, (d) contraste irregular ou artefatos na parede do oocisto, (e) oocistos mal focados, e (f) oocistos mal posicionados.

Dentre os problemas típicos que alteram a morfologia dos oocistos e que prejudicam o processo de reconhecimento, podemos citar os listados abaixo:

- **Oocistos achatados** (Figura 2.3(b)) – Conforme foi explicado na Seção 2.3.1, o oocisto sofre uma alteração no tamanho por causa da pressão da lamínula em função da pouca quantidade de líquido.
- **Oocistos encolhidos** (Figura 2.3(c)) – A parede do oocisto apresenta-se de forma irregular com a sensação de que o conteúdo aquoso interno se perdeu, causando um efeito visual de “encolhimento”. Isso pode ser causado por efeitos osmóticos e pode eventualmente indicar que os parasitas no seu interior estão inviáveis.

- **Oocistos com contraste irregular ou com artefatos na parede** (Figura 2.3(d)) – O oocisto não apresenta uma boa definição do contorno do oocisto ou então a linha de contorno é heterogênea. Esse artefato pode ter diferentes origens como a variabilidade natural do oocisto e a presença de debris aderidos à parede.
- **Oocistos mal focados** (Figura 2.3(e)) – Como os oocistos são estruturas bastante grandes, é praticamente impossível obter-se uma profundidade de campo que cubra toda a sua largura. É importante que o plano focal ajustado manualmente no microscópio priorize o contorno do oocisto, caso contrário a segmentação de sua imagem será prejudicada.
- **Oocistos mal posicionados** (Figura 2.3(f)) – Os oocistos exibem melhor as suas diferenças morfológicas quando estão na posição deitada, na qual os formatos circular, ovóide ou elíptico estão mais evidentes. Caso um oocisto seja fotografado em uma posição de pé, o seu contorno será sempre circular.
- **Oocistos com parede rompida ou deformada** – Nesse caso torna-se impossível definir o contorno do oocisto devido à falta de continuidade da parede. Esse artefato pode ocorrer devido a uma compressão muito forte ou abrasão durante a sua purificação.

Conforme exposto acima, os problemas apresentados impedem a automatização total da aquisição de imagens e detecção de objetos, uma vez que muitos dos parâmetros devem ser ajustados manualmente. Contudo, seguindo-se alguns critérios de qualidade, ilustrados na Figura 2.2, foi possível se padronizar as condições de captura e escolha adequada das imagens. Uma vez definidas essas condições, tornou-se uma tarefa relativamente fácil a segmentação e pré-processamento das imagens.

2.4 Pré-processamento de imagens

A rotina de pré-processamento geralmente consiste em uma série de operações que são aplicadas na imagem no nível mais baixo de abstração (intensidade da imagem). O objetivo do pré-processamento é a melhoria das imagens visando suprimir possíveis distorções ou destacar algumas características importantes da imagem para os processamentos subsequentes (Sonka et al., 1999).

No caso de micrografias de células ou microorganismos, são necessários procedimentos que resultem na detecção de objetos através da delimitação do contorno do objeto de interesse e seu posterior isolamento. O ideal seria realizar esses procedimentos automaticamente, mas devido à sua complexidade, essa tarefa requer o desenvolvimento de algoritmos robustos para detecção de

objetos. Um exemplo ilustrativo é reportado por Long et al. (2005), o qual realizou a detecção de células utilizando redes neurais.

Para o diagnóstico automático de espécies de *Eimeria*, modelo de estudo do presente trabalho, a etapa de pré-processamento tem como objetivo a detecção do contorno do oocisto. Após a etapa de recorte de oocistos individuais, tem-se os seguintes procedimentos: transformação das imagens em tons ou níveis de cinza, equalização, binarização e detecção do contorno (vide Figura 2.1).

2.4.1 Transformação em tons de cinza

As imagens originais dos oocistos isolados são capturadas em cores de 24 bits, mas, para efeito de processamento digital, optou-se por trabalhar em níveis de cinza. Essa escolha foi feita porque a coloração está sujeita a um número muito maior de variáveis como, por exemplo, a coloração do líquido de suspensão dos oocistos (geralmente uma solução de dicromato de potássio 2,5%), o tipo de lâmpada do microscópio, os filtros de luz usados no microscópio, a intensidade da lâmpada do microscópio, que pode mudar a temperatura da luz, entre outros.

Uma imagem digital colorida pode ser representada em distintos modelos de cores (Wyszecki e Stiles, 2000), sendo que pelo menos três matrizes 2D são necessárias para representar uma imagem colorida. Alguns dos mais populares sistemas de cores são: RGB (*Red, Green, Blue*), CMY (*Cyan, Magenta, Yellow*), e HSI (*Hue, Saturation, Intensity*). Por exemplo, num sistema como o RGB, a imagem pode ser definida como (Gonzales e Woods, 2002):

$$I_k = f_k(x, y) = f_k(\mathbf{x}) = \langle R_{\mathbf{x}}, G_{\mathbf{x}}, B_{\mathbf{x}} \rangle \quad (2.1)$$

Onde $\mathbf{x}$ é um vetor de coordenadas de um ponto na imagem I_k , e $f_k(\mathbf{x})$ é o nível das componentes R , G e B nesse ponto (*pixel*).

O padrão NTSC (International Telecommunication Union, 1998) estabeleceu uma formula para calcular o valor da intensidade do nível de cinza a partir dos componentes RGB:

$$\wp(x, y) = T[I_k] = T[f_k(\mathbf{x})] = 0.299R_{\mathbf{x}} + 0.587G_{\mathbf{x}} + 0.114B_{\mathbf{x}} \quad (2.2)$$

Um método simples e amplamente utilizado para gerar imagens em níveis de cinza consiste em realizar uma transformação da imagem colorida que calcule a média das componentes RGB de cada *pixel*. Isso pode ser descrito através da seguinte equação:

$$\wp(x, y) = T[I_k] = T[f_k(\mathbf{x})] = \frac{R_{\mathbf{x}} + G_{\mathbf{x}} + B_{\mathbf{x}}}{3} \quad (2.3)$$

Em nosso trabalho decidimos utilizar a Equação 2.3 para converter as imagens de cores de 24 *bits* para tons de cinza de 8 *bits*.

2.4.2 Equalização da imagem

Com o objetivo de contornar as limitações e variabilidades da iluminação no processo de aquisição de imagens digitais por microscopia, operações de equalização foram aplicadas sobre as imagens em tons de cinza.

O histograma pode ser usado como uma ferramenta para guiar os algoritmos de transformação de níveis de cinza para efeitos de filtragem. Uma transformada muito útil é a *equalização por histograma*, a qual permite o mapeamento dos níveis de cinza p em novos níveis de cinza q , de forma que a distribuição dos níveis de cinza q seja uniforme (Ballard e Brown, 1982).

Embora o resultado da equalização por histograma seja uma imagem com distribuição uniforme dos níveis de cinza, ela pode ser usada como uma forma de conseguir imagens menos sujeitas a variações da iluminação através de um processo de normalização de imagens (Drew et al., 1998). Esse procedimento também é conhecido como “casamento de histogramas” (*histogram matching* ou *histogram specification*) (Gonzales e Woods, 2002), o qual é uma generalização da equalização por histograma. O objetivo é gerar uma nova imagem com um histograma similar a um outro previamente especificado. Assim, uma imagem com problemas de iluminação pode ter seu histograma transformado, usando-se como referência o histograma de uma imagem padrão com iluminação adequada (“*eigenimage*”) (Huber et al., 2005).

A dificuldade do procedimento está em determinar a função de densidade de probabilidade que o histograma assume. Geralmente a função de densidade de probabilidade mais utilizada é a densidade Gaussiana (Gonzales e Woods, 2002), a qual é definida por dois parâmetros, a média μ e o desvio padrão σ (Equação 2.4).

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (2.4)$$

Na Figura 2.4 apresenta-se o diagrama do processo de equalização, onde a entrada está constituída por duas imagens, a imagem a ser equalizada (imagem original) e a imagem padrão (*eigenimage*). Por exemplo, pode-se verificar que o histograma da imagem original apresenta maior número

Figura 2.4: Equalização de imagem por “casamento” de histograma. A imagem original (a) foi capturada em condições de baixa luminosidade. Utilizando-se um histograma de uma imagem padrão (b), procedeu-se a uma transformação do histograma da imagem original de forma a “casar” com o histograma da imagem padrão, resultando assim numa imagem equalizada.

de tons de cinza escuros (mais para a esquerda do gráfico), enquanto que a imagem padrão apresenta tons de cinza mais claros (mais para a direita do gráfico). A imagem resultante (equalizada) apresenta tons mais claros, o que é confirmando ao verificar o seu histograma que está praticamente emparelhado com o histograma da imagem padrão. Essa conversão permite homogeneizar os histogramas de todas as imagens de um conjunto de treinamento, assim como as imagens de consulta, fazendo com que variações de luminosidade na captura, dentro de certos limites, sejam minimizadas.

2.4.3 Segmentação por limiarização

A limiarização (*thresholding*) é uma técnica simples e bem conhecida para a segmentação de imagens (Haralick e Shapiro, 1992; Sahoo et al., 1988). A limiarização é a operação de conversão de uma imagem de tons de cinza em uma outra imagem binária (preto e branca), processo também conhecido como “binarização”. Em uma imagem binária, cada valor de *pixel* é representado por um simples dígito binário. Na sua forma mais simples, a limiarização é uma operação que atribui o valor de 0 ou 1 a cada *pixel* da imagem baseado em uma comparação com um valor global de limiarização L .

$$f_L(x,y) = \begin{cases} 1 & \text{se } f(x,y) \geq L; \\ 0 & \text{se } f(x,y) < L. \end{cases} \quad (2.5)$$

A binarização é um passo adequado no pré-processamento porque permite uma redução significativa no armazenamento de informação e o tratamento da imagem binária conduz a uma análise mais simples. As imagens binárias permitem o uso de poderosas ferramentas de análise morfológica de formas (Costa e Cesar Jr., 2000). A Figura 2.5(b) apresenta uma imagem binarizada, obtida a partir de uma imagem de tons de cinza (Figura 2.5(a)) através da aplicação de um valor de limiarização de 130, considerando-se que a faixa dos níveis de cinza varia de 0 a 255.

Figura 2.5: Diferentes etapas da rotina de pré-processamento de uma image de oocisto: (a) imagem original colorida, (b) imagem em níveis de cinza, (c) imagem segmentada (binarizada) e (d) detecção do contorno.

2.4.4 Detecção do contorno

O resultado da etapa de binarização é utilizado para detectar o objeto de interesse na imagem, sobre o qual será realizada a detecção do contorno. Isto é feito através do algoritmo de seguimento de contorno (*contour following*), que consiste na extração paramétrica do contorno de uma imagem binária (para detalhes do algoritmo consultar (Costa e Cesar Jr., 2000)). O algoritmo inicia-se com a seleção de um ponto inicial que faz parte do contorno externo do objeto, assumindo-se que os *pixels* do objeto são pretos (*valor* = 0) e os *pixels* do fundo da imagem são brancos (*valor* = 1). Através de sucessivas chamadas recursivas, o objetivo é detectar o *pixel* seguinte do contorno, daí o termo “seguimento” de contorno. Para isso, faz-se uso de códigos de direção (*chain-code directions*) (Figura 2.6), e o resultado é uma representação paramétrica onde cada ponto do contorno é identificado por suas coordenadas $x(t)$ e $y(t)$.

Na Figura 2.5(c) é apresentado o resultado da detecção do contorno a partir de uma imagem binária (Figura 2.5(b)). A detecção do objeto e determinação do seu contorno são os primeiros passos no processo de caracterização dos oocistos, embora essa tarefa não seja totalmente automatizada

Figura 2.6: Processo de detecção de contorno através do algoritmo de seguimento de contorno (*contour following*). Os códigos de direção (*chain-code directions*) estão indicados.

devido à qualidade variável da imagem e/ou a presença de material indesejado (ruído) ao redor do objeto de interesse, o que a torna em um dos desafios ainda não resolvidos na área de visão artificial.

Capítulo 3

Representação de formas

3.1 Introdução

A idéia de representação de formas vem desde tempos remotos, Aristóteles (384–322 a.C.) manifestava que a mente é um “lugar das formas”, ou uma “forma das formas”. Existe também uma conexão mitológica ao deus grego dos sonhos, Morpheus. Os gregos acreditavam que as imagens mentais da vida real e dos sonhos procediam de uma mesma origem. A formalização da representação de formas começou com o extraordinário trabalho, *On Growth and Form*¹, onde D’Arcy Thompson (1942) estabeleceu as bases para a análise de formas morfológicas, idéias que ainda permanecem atuais. Thompson percebeu que as formas complexas são originadas a partir de princípios simples como, por exemplo, aspectos geométricos e topológicos da forma são expressos ao longo do desenvolvimento. Isso o levou a reinterpretar o desenvolvimento e a estrutura dos organismos em termos físicos e matemáticos. Esse foi um substancial avanço na quantificação de formas biológicas que, no entanto, teve pouca influência na época da publicação do trabalho. Naquela época, a biologia era vista principalmente em termos de anatomia comparativa junto com os princípios da teoria da evolução². Para Thompson, as mudanças na forma no decorrer do tempo (desenvolvimento) aconteciam principalmente pela ação de forças físicas, as quais eram as manifestações de vários tipos de energia.

Uma das idéias de Thompson indicava a noção de que mudanças temporais afetavam o organismo todo e não só alguns dos seus componentes. Isso levou Zuckerman (1950) a questionar se

¹Originalmente publicado em 1917.

²Thompson foi um morfologista solitário que rejeitava o Darwinismo (seleção natural) em favor de sua visão que proclamava que os organismos (devido a sua inerente plasticidade) poderiam se adaptar prontamente a novas restrições funcionais.

as formulações numéricas do tamanho e da forma poderiam ser derivadas a partir de algumas leis fundamentais da biologia, ou se alguns processos biológicos poderiam igualmente ser derivados, em algum sentido, de uma análise do tamanho e da forma. Esse é um dos desafios ainda não resolvidos pela comunidade científica.

Recentemente, Lestrel (2000) propôs o termo “morfometria estrutural” para a caracterização não só de aspectos geométricos, mas também da estrutura da superfície e da estrutura interna do organismo, onde a identificação da textura tem especial atenção. Por causa disso, destaca que é necessário lidar com técnicas de multi-escala, onde técnicas como análises por transformada de Fourier e *wavelets* têm-se mostrado adequadas. Nesse sentido, o aumento da velocidade de processamento dos computadores tem permitido o incremento da aplicabilidade dessas técnicas, as quais são bastante complexas, especialmente com o uso de imagens digitais (Zhang e Lu, 1974).

Do ponto de vista formal, as imagens podem ser entendidas em termos matemáticos como um conjunto de pontos conectados em um espaço F , o qual pode ser aproximado a um espaço binário discreto. A classificação de imagens, efetuada diretamente em F , torna-se um processo computacionalmente pesado que precisaria de $O(N^2)$ comparações, considerando-se que cada imagem está constituída de N *pixels*. A representação de uma imagem pode ser modificada mediante a aplicação de transformações de imagens, o que significa mapear o espaço original F para um novo espaço F' , tipicamente menor. Isto significa que grande parte da informação relacionada com a classificação é “reduzida” para um número relativamente menor de características, permitindo a redução da dimensão do espaço de características.

De fato, esse processo de transformação também está presente no sistema visual humano, que usando “conjuntos de filtros” pode extrair as características necessárias para reconhecer os detalhes que diferenciam um padrão de outro (Regan, 2002). Esses filtros possuem funcionalidades e sensibilidade próprias que, em conjunto, permitem representar o ambiente visual da maneira mais concreta. Há quem acredite que o sistema visual exista para derivar da imagem a informação que precisamos, e não simplesmente para recriar a imagem projetada na retina (Braddick et al., 1978). No caso da formação imagens 3D, Julesz (1995) reportou que o cérebro as constrói usando pequenas diferenças em cada imagem, o que o levou a inventar as imagens estereoscópicas para explicar sua teoria.

A idéia por trás das características resultantes da transformação é que determinadas operações de transformação permitam explorar e remover informação redundante, usualmente encontrada em imagens naturais (Kersten, 1987; Barlow, 1994; Olshausen e Field, 2000). Entretanto, deve-se observar que a compacidade³ não é a única característica que deve ser buscada em sistemas de visão compu-

³Em topologia, a compacidade é um conceito relacionado com a pequenez de um conjunto. De fato, qualquer conjunto finito é compacto.

tacional, mas também maximizar a capacidade da representação para salientar aspectos visuais mais relevantes, os quais têm alta incidência no sucesso dos algoritmos de classificação. Para isso, esses procedimentos devem caracterizar distintos aspectos da imagem, como bordas, cor, profundidade, textura e forma (Levine, 1985), luminosidade, movimento e disparidade binocular (Regan, 2000), entre outros.

A seguir serão detalhadas algumas das técnicas de visão computacional usadas neste trabalho, no intuito de transformar os objetos da imagem em uma representação mais simplificada que permita o tratamento e análise computacional. Para isso, as imagens são transformadas em um vetor de características, constituído por um grupo de valores que identificam três tipos de características morfológicas: (a) medidas geométricas, (b) caracterização da curvatura, e (c) quantificação da estrutura interna.

3.2 Medidas geométricas

Esta seção apresenta uma série de medidas simples, ou descritores gerais, muitos deles relacionados a aspectos métricos da forma. Essas medidas são úteis quando o tamanho da forma é importante, como é o nosso caso de aplicação, onde algumas espécies de oocistos podem ser facilmente diferenciadas considerando-se somente o seu tamanho, mas, no caso de outras, existe uma sobreposição, o que torna necessário se considerar outras características morfológicas. Os descritores gerais são medidas simples relacionadas com a medição da forma do objeto, entre as quais temos a área, diâmetros e simetria. Neste trabalho, aplicou-se a análise das componentes principais (Costa e Cesar Jr., 2000) do objeto com a finalidade de medir os diâmetros e o grau de simetria dos oocistos. Outros descritores gerais incluem a área (número de *pixels* que compõem o objeto), excentricidade (diâmetro maior/diâmetro menor), circularidade ($\text{perímetro}^2/\text{área}$) e a energia de dobramento (*bending energy*) (Young et al., 2004).

3.2.1 Área

A forma mais simples de se estimar a área de um objeto é contando o número de *pixels* que pertencem ao objeto. O contorno paramétrico (coordenadas dos *pixels*) do objeto, calculado previamente (Seção. 2.4.4), é traduzido para uma matriz binária, onde $\mathcal{B}(x,y) = 1$ representa um *pixel* pertencente ao contorno do objeto, e $\mathcal{B}(p,q) = 0$ representa um *pixel* que não é parte do contorno.

O procedimento que estima a área do objeto segue a lógica do algoritmo de preenchimento de

área de vizinhança 4-conectado que, a partir de um ponto de início (qualquer ponto dentro do objeto), começa a percorrer de forma recursiva todos os *pixels*, limitado pelo contorno paramétrico (Hearn e Baker, 1997).

3.2.2 Diâmetros

O diâmetro de um objeto é normalmente definido como a maior distância entre qualquer par de pontos pertencentes ao objeto. Um dos algoritmos que calcula o diâmetro é conhecido como de força bruta, que consiste em se buscar a distância máxima entre todos os pares de pontos que constituem o objeto. Embora esse algoritmo de força bruta não tenha maior complexidade para ser implementado, ele só auxilia no cálculo do diâmetro maior (comprimento). Entretanto, em formas biológicas é muito importante também se conhecer o diâmetro menor (largura) do objeto.

Uma abordagem importante, mediante a qual pode ser feito o cálculo dos diâmetros, refere-se ao conceito de autovalores. O cálculo dos diâmetros aplicando autovalores é composto de quatro passos: (1) determinação dos eixos principais, (2) translação do centro de massa do objeto à origem dos eixos cartesianos, (3) emparelhamento dos eixos principais do objeto com os eixos cartesianos, e (4) detecção da intersecção do contorno paramétrico com os eixos.

1. No intuito de entender o primeiro passo, isto é, a determinação dos eixos principais, considere-se a forma apresentada na Figura 3.1(a). A direção no sentido em que a forma é mais alongada (por exemplo, a direção ao longo da qual os pontos da forma são mais dispersos) é conhecida como “eixo maior”. Na Figura 3.1(a) o eixo maior está indicado como a maior reta dentro do objeto. A linha perpendicular ao eixo maior indica o “eixo menor”. Os eixos maior e menor são denominados como “eixos principais”. O cálculo dos eixos principais tem relação muito próxima com os autovetores das matrizes de covariância na teoria de probabilidade multivariada (Duda et al., 2001), mas, nessa situação, os vetores aleatórios correspondem aos componentes do contorno paramétrico ($x(t)$ e $y(t)$), previamente calculados na Seção 2.4.4.

O Algoritmo 1 descreve a sequência de passos para calcular os eixos principais de um objeto em função do contorno paramétrico. Sendo o contorno paramétrico constituído por dois vetores-elemento, o algoritmo cria uma matriz bi-dimensional X (passos 1 – 5) com as componentes do contorno paramétrico ($x(t)$ e $y(t)$). No passo 6 é calculada a matriz de covariância K da matriz X . Em seguida, são calculados os autovetores e autovalores da matriz de covariância (passos 7 e 8) para, finalmente, determinar como eixo maior o autovetor associado com o maior autovalor. O eixo menor é definido como o autovetor associado com o menor autovalor. É importante salientar que os programas científicos de computador, como MATLAB[®] ou

SCILAB, possuem funções pré-definidas que podem ser usadas para o cálculo dos autovetores e autovalores de matrizes. Da mesma forma, existem também bibliotecas de funções (por exemplo, *GNU Scientific Library*) para a programação em linguagem C++ ou Fortran.

```

Entrada:  $x(t), y(t)$ 
Saída:  $EixoMaior, EixoMenor$ 
1  $n = \text{longitude}(x);$ 
2 para  $i=1$  até  $n$  faça
3    $X[i, 1] = x[i];$ 
4    $X[i, 2] = y[i];$ 
5 fim
6  $K = \text{covariancia}(X);$ 
7  $\text{Autovetor}[] = \text{eigenvector}(K);$ 
8  $\text{Autovalor}[] = \text{eigenvalue}(K);$ 
9  $EixoMaior = \text{Autovetor}$  associado com o maior  $\text{Autovalor}$ ;
10  $EixoMenor = \text{Autovetor}$  associado com o menor  $\text{Autovalor}$ ;
11 retorna  $EixoMaior, EixoMenor$ 

```

Algoritmo 1: Cálculo de eixos principais. Adaptado de Costa e Cesar Jr. (2000)

2. O centro de massa do objeto é usado como ponto de intersecção dos eixos principais. Mediante esse ponto, realiza-se a translação do objeto à origem dos eixos cartesianos (veja Figura 3.1(b)), o que significa que o centro de massa converte-se no ponto de origem (0,0) dos eixos cartesianos (Figura 3.1(b)). Essa operação é feita pela subtração das coordenadas do centro de massa de todos os pontos do contorno paramétrico do objeto.
3. O terceiro passo consiste no alinhamento dos eixos principais com os eixos cartesianos (Figura 3.1(c)). Essa operação é feita multiplicando-se a inversa da matriz composta pelos autovetores com cada ponto do contorno paramétrico.
4. Finalmente, os diâmetros são calculados localizando-se a intersecção do contorno com os eixos cartesianos, o que pode ser feito percorrendo-se os eixos principais (cartesianos) do objeto, tendo como início o centro de massa (origem cartesiano) até atingir o contorno do objeto (Figura 3.1(c)).

O procedimento descrito no Algoritmo 1 é apropriado para objetos convexos (considerando-se que os parasitas, objetos deste estudo, apresentam uma forma arredondada e sem cavidades). Além disso, o algoritmo permite também poupar tempo de processamento e calcular, na sequência, as simetrias vertical e horizontal do objeto, o que é detalhado a seguir.

Figura 3.1: Cálculo dos diâmetros do objeto baseado nas componentes principais. (a) Objeto em posição original e seus componentes principais, (b) translação do objeto à origem baseado no centro de massa, e (c) rotação do objeto através do alinhamento das componentes principais com os eixos cartesianos, e posterior cálculo dos diâmetros.

3.2.3 Simetria

A simetria representa uma característica importante na diferenciação de formas. Dentre os distintos tipos de simetria, neste trabalho foi aplicada a simetria bilateral, que é considerada o primeiro caso de um conceito geométrico da simetria (Weyl, 1980). O cálculo da simetria é simplificado considerando-se o procedimento realizado para encontrar os eixos principais de um objeto (Seção 3.2.2), em que esse objeto é transladado à origem cartesiana e rotacionado em função dos autovetores. Por exemplo, o objeto original (Figura 3.1(a)) resulta em uma imagem como mostrada na Figura 3.2(a), sobre a qual realiza-se a rotação em função do eixo das ordenadas – simetria em função do eixo maior (Figura 3.2(b)), e uma outra rotação em função do eixo das abscissas – simetria em função do eixo menor (Figura 3.2(c)).

O cálculo do nível de simetria é feito a partir de uma imagem binária. A Figura 3.2(d), por exemplo, é uma imagem binária na qual os elementos que fazem parte do objeto tem valor 1 (região branca), e o restante corresponde ao fundo da imagem com valor 0 (região preta). O primeiro passo do cálculo consiste em se refletir a forma em relação à linha que tem como orientação o eixo maior e que passa pelo centro de massa. A imagem resultante da reflexão pode apresentar buracos que devem ser preenchidos usando-se o operador de fechadura da morfologia matemática (Costa e Cesar Jr., 2000). A versão refletida é sobreposta na imagem original (adição de duas imagens), resultando em uma imagem em tons de cinza (Figura 3.2(e)). Nessa figura, os elementos da imagem apresentam três tipos de valores: 0 (fundo da imagem), 1 (porção da forma que é assimétrica) e 2

(*pixels* simétricos). Seja N_1 o número de *pixels* das regiões assimétricas (*pixels* com valor 1), e N_2 o número de *pixels* da região simétrica (*pixels* com valor 2), então o grau de simetria do objeto pode ser estimado usando-se a relação $N_2/(N_1 + N_2)$ (Costa e Cesar Jr., 2000). O mesmo procedimento é seguido quando a simetria é realizada em função do eixo menor (Figura 3.2(f)).

Figura 3.2: Cálculo da simetria baseado nas componentes principais. Depois que as componentes tenham sido alinhadas com os eixos cartesianos (a), o objeto é rotacionado em função do eixo maior (b) e do eixo menor (c). Os cálculos são feitos sobre a imagem binária (d), a partir da qual são produzidas outras imagens de simetria no eixo maior (e) e no eixo menor (f), onde a região branca representa a porção não simétrica do objeto.

3.3 Análise da forma por curvatura

A curvatura é uma das mais importantes características que podem ser extraídas do contorno da imagem. O trabalho de Attneave (1954) influenciou pesquisas posteriores na área de reconhecimento de forma por visão computacional, e enfatizou a importância que eventos transientes (não repetitivos) e assimetrias têm na percepção visual de humanos. Fortes evidências biológicas têm sido identificadas para o estudo da curvatura, que é aparentemente um importante aspecto explorado pelo sistema de visão humano. Nessa direção, Attneave (1954) define que o sistema visual não está baseado em *arrays* retilíneos, mas que a curvatura é uma característica específica da percepção da visão humana, cujos sensores encontram-se no nível neural.

Para o cálculo da curvatura, precisa-se do contorno paramétrico representado pelos sinais $x(t)$ e $y(t)$. Seja $c(t)$ a representação paramétrica do contorno:

$$c(t) = (x(t), y(t)) \quad (3.1)$$

a curvatura $k(t)$ de $c(t)$ é definida como:

$$k(t) = \frac{\dot{x}(t)\ddot{y}(t) - \ddot{x}(t)\dot{y}(t)}{(\dot{x}(t)^2 + \dot{y}(t)^2)^{3/2}} \quad (3.2)$$

Para o cálculo da curvatura, precisa-se estimar a primeira e segunda derivadas dos sinais $x(t)$ e $y(t)$. Uma das formas de se calcular a derivada é utilizando a propriedade derivativa de Fourier (Brigham, 1988). Sejam $X(f)$ e $Y(f)$ as transformadas de Fourier de $x(t)$ e $y(t)$ respectivamente, a propriedade derivativa de Fourier implica que:

$$\dot{X}(f) = j2\pi fX(f) \quad (3.3)$$

$$\dot{Y}(f) = j2\pi fY(f) \quad (3.4)$$

$$\ddot{X}(f) = -(2\pi f)^2 X(f) \quad (3.5)$$

$$\ddot{Y}(f) = -(2\pi f)^2 Y(f) \quad (3.6)$$

onde j é um número complexo e $\dot{X}(f)$, $\dot{Y}(f)$, $\ddot{X}(f)$ e $\ddot{Y}(f)$ denotam as transformadas de Fourier de $\dot{x}(t)$, $\dot{y}(t)$, $\ddot{x}(t)$ e $\ddot{y}(t)$, respectivamente. Assim, a aplicação da propriedade anterior de Fourier, seguida pela transformada inversa de Fourier, permite o cálculo da curvatura aplicando-se a Equação 3.2 em termos das transformadas de Fourier dos sinais $x(t)$ e $y(t)$.

3.3.1 Curvatura baseada na transformada multiescala de Fourier

A abordagem multi-escala baseada na transformada de Fourier para a estimação da curvatura foi proposta por Cesar Jr. e Costa (1995, 1996, 1997). O processo multi-escala para a estimação da curvatura gera os denominados curvogramas, onde os valores de curvatura são representados em uma dimensão escala-espço. Nesta abordagem, existem dois problemas principais: a carência de uma representação analítica do contorno original (a partir da qual a curvatura poderia ser calculada explicitamente), e a necessidade de estimação de derivadas numéricas. O primeiro problema pode ser contornado com a transformada de Fourier do contorno, através da aplicação da propriedade derivativa. Já o segundo problema é tratado através de filtragem passa-baixa do sinal, o que permite a generalização do método para uma abordagem multi-escala.

Um aspecto interessante da aplicação da propriedade derivativa da transformada de Fourier, é que ela permite calcular a derivada e a integral de uma função, onde a ordem da derivação/integração é um número real. Seja o par genérico de Fourier $q(t) \leftrightarrow Q(f)$ e seja a um real não negativo, que define a ordem da derivação, a propriedade derivativa de Fourier é definida por:

$$\frac{d^a q(t)}{dt^a} = \mathfrak{F}^{-1}\{D_a(f)Q(f)\} \quad (3.7)$$

Onde $D_a(f) = (j2\pi f)^a$ e j é um número imaginário. Para calcular a integral da função, deve-se tomar $a < 0$ (Costa et al., 2001).

A aplicação da propriedade acima para funções discretas acaba gerando uma derivada com muito ruído (efeito “*shrinking*”). A solução nesse caso é aplicar um filtro Gaussiano (filtragem passa-baixas) na função. Isso é feito pela inclusão do termo $G(f, \sigma)$ (uma Gaussiana de desvio padrão σ definida no espaço das frequências f):

$$\frac{d^a q(t)}{dt^a} = \mathfrak{F}^{-1}\{D_a(f)Q(f)G(f, \sigma)\} \quad (3.8)$$

A função Gaussiana tem a característica de poder ter sua largura de banda controlada mediante o parâmetro σ , e a variação desse parâmetro permite obter diferentes curvas para a derivada (o que é denominado multi-escala (Cesar Jr. e Costa, 1996)). Isso significa que o sinal pode ser convoluído com um conjunto de Gaussianas, graças ao teorema da convolução de Fourier e ao fato de a transformada de Fourier de uma Gaussiana resulta ser uma outra gaussiana, isto é:

$$g(t, \sigma) = \exp\{-2\pi^2 t^2 \sigma^2\} \Leftrightarrow G(f, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{f^2}{2\sigma^2}\right\} \quad (3.9)$$

Assim, o cálculo da curvatura multi-escala, a partir do contorno paramétrico, consiste em se multiplicar sucessivas Gaussianas, que mudam seu parâmetro de desvio padrão (σ), com os espectros dos vetores x e y . Os espectros suavizados resultantes são denotados por $X_s(f)$ e $Y_s(f)$, e definidos como:

$$X_s(f) = X(f)G(f, \sigma) \quad (3.10)$$

$$Y_s(f) = Y(f)G(f, \sigma) \quad (3.11)$$

Como exemplo ilustrativo, considere-se o oocisto binarizado da Figura 2.5(b), que foi submetido a um seguimento de seu contorno, resultando em uma representação paramétrica (Figura 2.5(c)). O cálculo da curvatura começa seu percurso, ao redor do objeto, no ponto 1, movendo-se em sentido anti-horário, conforme ilustrado na Figura 3.3(a). O valor absoluto da curvatura ao longo do contorno do objeto, com $\sigma = 10$, é mostrado na Figura 3.3(b). Naturalmente, as partes mais encurvadas do objeto implicam maior valor, sendo essa a razão do valor absoluto da curvatura na Figura 3.3(b) apresentar “picos” ou “elevações” onde é maior a curvatura do objeto em questão. Da mesma forma, o resultado de aplicar o mesmo processo com $\sigma = 50$ (Figura 3.3(c)) resulta em uma curvatura mais suavizada. Sucessivos acréscimos no valor do parâmetro fazem com que a curvatura fique mais suave, o que pode ser visualizado no curvograma 3D da Figura 3.3(d).

Outros exemplos de curvatura são apresentados na Figura 3.4, onde os curvogramas foram gerados com $\sigma = 10$. O nível de detalhamento é menor quando a imagem é relativamente pequena (como na espécie *E. acervulina*, Figuras 3.4(a) e (b)), enquanto que o curvograma apresenta mais detalhes quando o objeto é maior (como na espécie *E. maxima*, Figuras 3.4(c) e (d)). Isto ocorre porque, dada uma resolução de captura, se o objeto for maior, seu contorno será definido por um número maior de *pixels* do que em um objeto menor.

Embora a curvatura possa ser utilizada como um vetor de características, isto pode apresentar sérias desvantagens, uma vez que o sinal da curvatura pode ser muito grande e altamente redundante. Uma vez que a curvatura tenha sido estimada, as seguintes medidas podem ser calculadas para se contornar esse problema (Costa et al., 2001):

- **Curvatura amostrada:** Ao invés de usar o total dos valores da curvatura ao longo do contorno, é possível amostrar o sinal da curvatura com o intuito de obter um conjunto de características pequeno.
- **Estatísticas da curvatura:** O histograma da curvatura pode fornecer uma serie de medidas

Figura 3.3: Contorno paramétrico de um oocisto (a) e seu correspondente curvograma usando $\sigma = 10$ (b) e $\sigma = 50$ (c), ou sucessivas variações do desvio padrão na função Gaussiana, mostrada em um curvograma 3D (d).

globais muito úteis, tais como a média da curvatura, mediana, variância, desvio padrão, entropia, momentos, entre outros.

- **Ponto máximo e mínimo de inflexão:** O fato de que nem todos os pontos ao longo do contorno são igualmente informativos implica que a análise pode ser restringida aos pontos dominantes, como aquela onde a curvatura é um máximo positivo ou um mínimo negativo, ou ponto de inflexão.
- **Energia de dobramento:** Esta medida tem uma fundamentação da Física, a partir da teoria de elasticidade. Essa medida significa a quantidade de energia requerida para transformar um

Figura 3.4: Contornos paramétricos de *E. acervulina* (a) e *E. maxima* (c) e seus respectivos curvogramas (b) e (d) usando $\sigma = 10$.

contorno fechado em uma circunferência de mesmo perímetro. Em consequência, esse descritor do contorno é invariante à translação, rotação e reflexão, sendo facilmente normalizado em relação a mudanças na escala e frequentemente utilizado como uma medida complexa. A média da energia de dobramento está definida, no caso discreto, como a soma dos quadrados dos valores de curvatura ao longo do contorno, dividido pelo número de pontos.

3.4 Caracterização da estrutura interna

3.4.1 Características para extração de textura

A textura de uma região da imagem esta determinada pela forma como os tons de cinza estão distribuídos sobre os *pixels* nessa região. Embora não tenha uma clara definição da textura, toda pessoa possui a capacidade de descrever uma imagem pela forma como se apresenta: fina, suave ou irregular, homogênea ou heterogênea, e assim por diante. Um dos objetivos do nosso trabalho foi o de estabelecer métodos que gerassem as características apropriadas para quantificar a estrutura interna do oocisto. Estas características exploram a relação de espaço subjacente na distribuição dos tons de cinza.

Os métodos de análise de textura têm sido classificados em quatro categorias (Tuceryan e Jain, 1998): estatísticos, geométricos, baseados em modelos e baseados em processamento de sinais. Um dos métodos mais usados é o que envolve as denominadas matrizes de co-ocorrência, popularizadas devido ao trabalho de Haralick et al. (1973). Esse método fornece uma abordagem de segunda ordem para gerar as características de textura. Embora sejam principalmente aplicadas à caracterização de imagens por textura, as matrizes de co-ocorrência também têm sido usadas na segmentação de regiões (Jobanputra e Clausi, 2006).

3.4.2 Matrizes de co-ocorrência

As matrizes de co-ocorrência fornecem a informação das posições relativas dos distintos níveis de cinza dentro da imagem. Isso implica no uso das probabilidades condicionais conjunta, C_{ij} , de cada combinação de par de pontos dos níveis de cinza. Essa combinação leva em consideração o vetor de deslocamento (dx, dy) , que representa a separação de um par de *pixels* nas direções x e y , respectivamente. Tradicionalmente, as probabilidades são representadas em uma matriz de co-ocorrência de níveis de cinza (*grey level co-occurrence matrix* – GLCM) (Hall-Beyer, 2005), onde o índice (i, j) na matriz representa a probabilidade C_{ij} :

$$C_{ij} = \frac{P_{ij}}{\sum_{i,j=0}^{g-1} P_{ij}}, \quad (3.12)$$

onde P_{ij} identifica a frequência de ocorrência entre dois níveis de cinza, i e j , para um determinado vetor de deslocamento (dx, dy) . g é o número de níveis de cinza quantificados.

Também é possível gerar a matriz de co-ocorrência baseada em dois parâmetros: (1) a distância

relativa entre os *pixels* e (2) sua orientação relativa. Seja d a distância relativa medida em número de *pixels* (por exemplo, $d = 1$ para *pixels* de vizinhança 8), a orientação é quantificada em quatro direções: horizontal, diagonal, vertical, e anti-diagonal (0° , 45° , 90° , 135°). A Figura 3.5 mostra a representação gráfica bi-dimensional da matriz de co-ocorrência da respectiva imagem dos oocistos.

Figura 3.5: Micrografias (imagens superiores) de oocistos de *E. mitis* (a), *E. brunetti* (b) *E. maxima* (c) e suas respectivas matrizes de co-ocorrência (imagens inferiores) geradas com $d = 2$ e orientação de 90° .

Distintas informações podem ser extraídas a partir da matriz de co-ocorrência, que na verdade representa um histograma de segunda ordem. Algumas dessas informações têm interpretação física direta em relação à textura como, por exemplo, para se quantificar a rugosidade, suavidade, entre outras. Por outro lado, algumas características não possuem tais propriedades, mas guardam alguma informação relacionada com a textura com um alto poder de discriminação (Theodoridis e Koutroumbas, 1998). Assim, por exemplo:

- **Segundo Momento Angular (SMA):** Esta característica é uma medida da suavidade da imagem. Quanto menos suave é a região, mais uniforme é a distribuição dos pontos $P(i, j)$ e menor o SMA.

$$SMA = \sum_{i,j=0}^{g-1} C_{ij}^2 \quad (3.13)$$

- **Contraste (CON):** Mede o contraste da imagem, assumindo valores maiores quanto mais contraste apresenta a imagem.

$$CON = \sum_{i,j=0}^{g-1} C_{ij}(i-j)^2 \quad (3.14)$$

- **Momento da Diferença Inversa (MDI):** esta característica assume valores maiores para imagens de baixo contraste.

$$MDI = \sum_{i,j=0}^{g-1} \frac{C_{ij}}{1 + (i-j)^2} \quad (3.15)$$

- **Entropia (ENT):** é uma medida da aleatoriedade e assume valores menores para imagens suaves.

$$ENT = - \sum_{i,j=0}^{g-1} C_{ij} \log C_{ij} \quad (3.16)$$

3.5 Espaço de características

Um método de extração de características cria um novo espaço como resultado de transformações ou combinações das características do espaço original. Formalmente, dado um espaço de características F , de dimensão N , um método de extração de características H é uma função $H : F \longrightarrow F'$, onde F' possui dimensão m , sendo que $m < N$.

Considerando os métodos anteriormente descritos nesta seção, a função H representaria o modelo de caracterização morfológica proposto neste trabalho, o mesmo que esta constituido por três tipos de características (a) medidas geométricas, (b) caracterização da curvatura, e (c) quantificação da estrutura interna. O conjunto de características extraídas forma um espaço 14-dimensional:

- **Características de ordem geométricas**
 - Diâmetro maior
 - Diâmetro menor
 - Simetria em função do diâmetro maior
 - Simetria em função do diâmetro menor
 - Área
 - Entropia da estrutura interna

- Excentricidade (altura/largura)
- **Características da curvatura**
 - Média da curvatura
 - Desvio padrão da curvatura
 - Entropia da curvatura
- **Características da textura**
 - Segundo momento angular
 - Contraste
 - Momento da diferença inversa
 - Entropia

Capítulo 4

Classificação e mineração

4.1 Introdução

O processo de classificar implica em designar classes ou categorias aos itens de acordo com as suas características. Existem particularmente dois aspectos importantes relacionados à classificação de padrões. O primeiro é, dada uma imagem de entrada, decidir se ela pertence a alguma classe específica previamente definida. Esta categoria de reconhecimento de padrões é usualmente conhecida como classificação supervisionada. O segundo aspecto, igualmente importante na classificação de padrões, consiste em se definir ou identificar as classes envolvidas numa população de padrões previamente não classificados. Esta é uma tarefa complexa que geralmente requer um aprendizado, sendo conhecida como classificação não supervisionada, aglomeração ou *clustering* (Hastie et al., 2001).

Um problema fundamental em Biologia é a classificação dos organismos, a qual foi classicamente realizada através de discriminação morfológica e, mais recentemente, incluiu marcadores moleculares. Uma vez que se dispõe de uma nova amostra biológica, o problema consiste em se classificá-la dentro das divisões taxonômicas. No presente trabalho foram utilizadas amostras de oocistos derivados de várias espécies do gênero *Eimeria*, todas elas devidamente classificadas do ponto de vista taxonômico, e com grau de pureza confirmado por inspeção visual e por ensaios de PCR. Assim, a aquisição de imagens foi feita com amostras puras, previamente identificadas, gerando-se, assim, um conjunto de características conhecidas para cada espécie. Para a classificação supervisionada, desenvolveu-se o modelo matemático que considera como entrada uma imagem de consulta e atribui-a a uma das classes das espécies previamente identificadas.

O processo de mineração de dados aplicado neste trabalho teve por objetivo a generalização do

modelo proposto, o que foi feito explorando o conjunto de dados em busca de padrões e relações entre as variáveis que permitiram determinar a dimensão adequada do espaço de características (seleção de características), definir o tamanho mínimo do conjunto treinamento e avaliar o desempenho dos classificadores.

A ferramenta utilizada para o reconhecimento supervisionado foi um classificador Bayesiano (Box e Tiao, 1973; Duda et al., 2001; Theodoridis e Koutroumbas, 1998), o qual é adequado para este tipo de tarefa e é fundamentado em conceitos sólidos de Estatística (Pernkopf, 2005). Foram usadas duas abordagens de classificação Bayesiana, uma primeira por densidade normal (Seção 4.2.1) e a segunda por métodos não paramétricos (Seção 4.2.2). Cada classificador foi usado como função-critério para os processo de seleção de características e determinação do tamanho mínimo do conjunto de treinamento. Para a avaliação dos classificadores, além da taxa média de acerto (acurácia), também foram usadas curvas ROC para análise do seu desempenho (Seção 4.4.5).

4.2 Classificação Bayesiana

Dado um “vetor de características” $\mathbf{x}$ e c “classes”, $\omega_1, \omega_2, \dots, \omega_c$, formam-se c “probabilidades condicionais” $P(\omega_i|\mathbf{x})$, $i = 1, 2, \dots, c$, também conhecidas como “probabilidades a posteriori”. Cada uma delas representa a probabilidade de que o padrão desconhecido pertença à respectiva classe ω_i , considerando-se que o correspondente vetor de características adquira o valor $\mathbf{x}$. O objetivo do classificador Bayesiano é calcular o máximo desses c valores usando o teorema de Bayes (Box e Tiao, 1973; Duda et al., 2001; Theodoridis e Koutroumbas, 1998):

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}. \quad (4.1)$$

A fórmula de Bayes mostra que com o valor observado de $\mathbf{x}$ é possível se converter a probabilidade a priori $P(\omega_j)$ em uma probabilidade a posteriori $P(\omega_j|\mathbf{x})$. Denomina-se $p(\mathbf{x}|\omega_j)$ como a verossimilhança de ω_j em relação a $\mathbf{x}$. O fator de evidência, $p(\mathbf{x})$, pode ser visto meramente como um fator de escala que garante que a soma das probabilidades a posteriori seja 1.

Para minimizar a taxa de erro, deve-se selecionar o i que maximiza a probabilidade a posteriori $P(\omega_i|\mathbf{x})$. Em outras palavras, para se determinar a taxa mínima de erro:

$$\text{Decidir por } \omega_i \text{ se } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \text{ para todo } j \neq i. \quad (4.2)$$

Uma dificuldade para o uso do classificador Bayesiano é definir o modelo a ser utilizado como função de densidade (verossimilhança). Uma solução consiste no uso de múltiplos modelos (Gelfand et al., 1992; Gelfand e Dey, 1994) o que implicaria calcular uma média de modelos (Carlin e Louis, 1996) ou *ranking* de modelos (Madigan e Raftery, 1994). Em geral, o processo de seleção do modelo é arbitrário na tentativa de eleger um modelo que forneça uma explicação confiável dos dados. Entre as funções de densidade, podemos citar a normal, Dirichlet, lognormal, gamma, Poisson, entre outras.

4.2.1 Classificação por densidade normal

Embora a regra de decisão Bayesiana não seja uma função discriminante, ela equivale a particionar o espaço de características em c regiões, para um problema de c classes. Se as regiões R_i e R_j são contíguas, isso significa que elas estão separadas por uma superfície de decisão no espaço de características multidimensional. Essas separações podem ser definidas em termos de funções discriminantes $\mathbf{g}_i(\mathbf{x})$, onde:

$$\mathbf{g}_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)}, \quad (4.3)$$

$$\mathbf{g}_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i). \quad (4.4)$$

Uma das mais comuns funções de verossimilhança encontradas na prática é a Gaussiana, ou função de densidade normal. Uma das maiores razões da sua popularidade está relacionada ao seu fácil tratamento computacional e ao fato de ela modelar adequadamente uma grande quantidade de casos, especialmente aqueles ligados a fenômenos naturais (Huxley, 1993). A densidade normal é completamente determinada pelos valores numéricos de dois parâmetros, a média μ e a variância σ^2 . Quando se trabalha com um alto número de variáveis ela é denominada *densidade normal multivariada*:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left[-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu) \right], \quad (4.5)$$

onde $\mathbf{x}$ é um vetor coluna d -dimensional, μ é o vetor média d -dimensional, Σ é a matriz de covariância de dimensão $d \times d$, e $|\Sigma|$ e Σ^{-1} são sua determinante e inversa, respectivamente. Verifica-se que $(\mathbf{x} - \mu)^t$ corresponde à transposta de $\mathbf{x} - \mu$.

Finalmente, utilizando-se as normais multivariadas (Equação 4.5) dentro das funções discriminantes (Equação 4.4), isto é, se $p(\mathbf{x} | \omega_i) \sim N(\mu_i, \Sigma_i)$, as discriminantes multivariadas resultam na seguinte equação:

$$\mathbf{g}_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^t \Sigma_i^{-1}(\mathbf{x} - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i). \quad (4.6)$$

Esta equação é completamente definida somente por dois parâmetros: o vetor média μ_i e a matriz de covariância Σ_i .

Similaridade de imagens

Depois de o elemento $\mathbf{x}$ ter sido designado a uma classe ω_i usando-se o classificador Bayesiano com função de verossimilhança normal, o passo seguinte consiste em se determinar o nível de similaridade entre a imagem de consulta e a espécie designada. Nesse sentido, é preciso se calcular o elemento protótipo de cada classe, o qual corresponde ao vetor média μ da densidade normal. Considerando-se um conjunto de treinamento constituído pelos exemplos $\mathbf{x}_1, \dots, \mathbf{x}_n$, o protótipo deste conjunto será o vetor média dos exemplos. Assim, adotou-se este protótipo como o elemento mais representativo para cada classe.

A distância de Mahalanobis (Duda et al., 2001; Theodoridis e Koutroumbas, 1998) pode ser usada como medida de similaridade entre o elemento $\mathbf{x}$, classificado na classe ω_i , e o seu protótipo μ_i correspondente. Esta distância é adequada para dados multivariados normais, que como característica estão agrupados ao redor do vetor média μ , formando uma nuvem de formato elipsoidal cujos eixos principais são os autovetores da matriz de covariância Σ . Assim, a medida natural da distância de $\mathbf{x}$ para a média μ é dada por:

$$r^2 = (\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu). \quad (4.7)$$

Por este motivo, o classificador Bayesiano que usa como função de verossimilhança a normal, foi denominado neste trabalho como classificador por similaridade.

4.2.2 Classificação por métodos não paramétricos

Algumas dificuldades podem se apresentar com os métodos paramétricos como a distribuição normal. Por exemplo, se houver um acréscimo no desvio padrão da classe ω_i ($\sigma_{\omega_i}^2$) e um incremento

no número de classes (c), isso irá afetar sensivelmente a forma da função de densidade, o que implica que os parâmetros desse modelo não serão totalmente identificados a partir dos dados. Alguns desses problemas e possíveis soluções são discutidos por West (1992) e Diebolt e Robert (1994).

A alternativa aos modelos paramétricos é o uso de procedimentos não paramétricos de análise de densidade. Isto implica que a distribuição a ser usada seja indicada a partir dos dados (Salinas-Torres et al., 2002; Carlin e Louis, 1996). Nesse sentido, Pereira e Stern (1999, 2001) introduziram o teste de significância genuinamente Bayesiano (FBST - *Full Bayesian Significance Test*), o qual utiliza apenas a distribuição a posteriori na avaliação da evidência em favor da hipótese. Esse teste embute o popular *processo a priori Dirichlet*, introduzido por Ferguson (1973), o qual é uma generalização multivariada da distribuição beta.

A função de densidade de probabilidade da distribuição Dirichlet de ordem K é um vetor K -dimensional com um conjunto de vetores $x = (x_1, \dots, x_K)$ e com $x_i \geq 0$:

$$f(x; \alpha) \sim \prod_{i=1}^K x_i^{\alpha_i - 1} \delta \left(1 - \sum_{i=1}^K x_i \right) \quad (4.8)$$

onde $\alpha = (\alpha_1, \dots, \alpha_K)$ é um vetor parâmetro com $\alpha_i \geq 0$. O delta Dirac δ assegura que a densidade não seja zero

$$\sum_{i=1}^K x_i = 1. \quad (4.9)$$

A constante de normalização é a função beta multinomial, a qual é expressa em termos da função gamma:

$$\frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} = B(\alpha). \quad (4.10)$$

Então, a densidade pode ser escrita como a função:

$$g(x; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1} \quad (4.11)$$

tendo como domínio o conjunto de vetores x com K -componentes reais não negativos com $|x|_1 = 1$.

A classificação dos elementos é feita a partir da matriz de frequências amostrais de cada característica, onde as linhas representam as classes e as colunas representam os cortes de classificação

(definidas pelos percentis). Com a matriz de frequência, aplica-se a distribuição de Dirichlet de ordem C (número de classes) com vetor da distribuição a priori $(\pi_1, \dots, \pi_C)$, onde $\sum_{i=1}^C \pi_i = 2$. O resultado dessa operação é uma matriz $C \times C$ que representa a distribuição conjunta de classificação e classe. Com esses dados, pode-se obter as probabilidades a posteriori de um elemento pertencer a uma classe tendo recebido uma determinada classificação (Pereira, 2006).

4.3 Processo de classificação

O processo de diferenciação consiste na classificação de cada imagem em uma das distintas espécies de oocistos. Para isso, é preciso que o classificador seja modelado de acordo com o domínio do conjunto de dados, o que se denomina como processo de aprendizado. Neste trabalho, foi utilizado o aprendizado supervisionado, que consiste na provisão de “dados de treinamento” (elementos com categorias conhecidas), com a finalidade de se definir os parâmetros do modelo de classificação. Para se avaliar o classificador, é necessário um outro grupo de exemplos, os “dados de teste”, os quais são tratados pelo classificador como elementos anônimos. Isto permite conhecer o grau de acurácia do classificador, comparando-se a classe original de cada elemento do conjunto de teste com a classe designada pelo classificador.

Figura 4.1: Fluxo de trabalho do processo de classificação utilizado neste trabalho.

Na Figura 4.1 apresenta-se a seqüência do processo de classificação supervisionada utilizado neste trabalho. O conjunto de dados é separado em dois subconjuntos (conjunto de treino e conjunto de teste), sendo o primeiro utilizado na geração do modelo de classificação (etapa conhecida como “treinamento”), e o segundo subconjunto usado na avaliação do classificador treinado. Esse procedimento pode ser realizado várias vezes (retroalimentação). No final do processo, obtém-se uma

matriz de classificação (ou matriz de confusão), onde cada elemento da matriz indica o número de elementos da classe C_i classificados na classe C_j ($i, j = 1, 2, \dots, n$ onde $n = \text{número de classes}$). A diagonal da matriz de classificação indica os elementos corretamente classificados.

A separação do conjunto de dados (conjunto de treino e conjunto de teste) é feita com o intuito de se analisar o comportamento do classificador perante distintas situações. Neste trabalho, a separação foi feita de duas formas:

1. *Leave-one-out* ou extração de elementos individuais. Nesse procedimento, um elemento do conjunto de dados é extraído para ser usado como elemento de teste (conjunto de teste) e todos os demais elementos são usados como conjunto de treino. Essa abordagem é reiterada devolvendo-se o elemento testado na etapa anterior e escolhendo-se um novo para ser usado como teste. A vantagem desse método é que cada elemento é confrontado com todo o restante, permitindo assim que o classificador seja treinado com o máximo número de elementos e, portanto, haja uma perda mínima de informação. O problema com esta abordagem é que se o conjunto de dados for muito grande, o processamento requerido será extremamente intenso.
2. Separação de subconjuntos aleatórios, para o qual é preciso se definir a porcentagem de elementos a serem usados como conjunto de treino. Com esse dado, seleciona-se aleatoriamente os elementos que farão parte do conjunto de treino. O restante é usado como conjunto de teste. Por exemplo, se a porcentagem de elementos a ser usada como treino é 30%, serão escolhidos de forma aleatória os 30% de elementos de cada classe, deixando o restante, 70%, como conjunto de teste. A desvantagem dessa abordagem é de que o subconjunto de treino escolhido pode estar viciado, por causa da perda de informação. Por esse motivo recomenda-se realizar o procedimento reiteradamente, e calcular o resultado final a partir da média dos resultados parciais, procurando, assim, reduzir a potencial perda de informação.

O Algoritmo 2 descreve em pseudocódigo o modelo matemático/estatístico de classificação supervisionada (função CLASSIFICA), que usa como discriminador o classificador Bayesiano.

O Algoritmo 2 requer um conjunto de dados (*Dados*) com um número definido (N_c) de classes e um número (N_f) de características. A partição é definida pela proporção $\%treino : \%teste$, sendo o número de iterações do processo aleatório de partição determinado pelo parâmetro *NrandomParticiona*. Adicionalmente, o parâmetro *LC* define o número de ciclos de aprendizado do classificador. A matriz de classificação resultante é *MclassMedia*.

Para um melhor entendimento do algoritmo descrito acima, os processos de particionamento e de classificação estão representados como implementações separadas. A função PARTICIONA é

```

Entrada: Dados;
Entrada:  $N_c \leftarrow$  # de classes;
Entrada:  $N_f \leftarrow$  # de características;
Entrada:  $\%treino \leftarrow$  % do conjunto de treino;
Entrada:  $\%teste \leftarrow$  % do conjunto de teste;
Entrada:  $N_{randomParticiona} \leftarrow$  # de conjuntos aleatórios;
Entrada:  $Ciclos \leftarrow$  # de ciclos de aprendizado;
Saída:  $M_{classMedia}[][]$ 
1 inicializa  $M_{classAux}[][]$  com zeros;
2 para  $i=1$  até  $N_{randomParticiona}$  faça
3    $[ConjTreino, ConjTeste] = PARTICIONA(Dados, \%treino, \%teste, N_c);$ 
4    $M_{class} = CLASSIFICABAYES(ConjTreino, ConjTeste, N_c, N_f, CA);$ 
5    $M_{classAux} = M_{classAux} + M_{class};$ 
6 fim
7  $M_{classMedia} = M_{classAux} / N_{randomParticiona};$ 
8 retorna  $M_{classMedia};$ 

```

Algoritmo 2: Processo de classificação

a responsável pelo processo de particionamento aleatório do conjunto de dados (*Dados*), usando os seguintes parâmetros de entrada: o conjunto de dados, a proporção treino:teste, e o número de classes. A função *PARTICIONA* retorna, então, como resultado, os respectivos subconjuntos de treino e teste. A função *CLASSIFICABAYES* é o processo principal que implementa o classificador. O classificador é treinado com o subconjunto *ConjTreino* e avaliado pelo subconjunto *ConjTeste*. Ambas as tarefas também usam como entrada o número de classes, de características, e de ciclos de treino. Então, a função *CLASSIFICABAYES* retorna a matriz de confusão *Mclass*. Finalmente, *MclassMedia* é a matriz de confusão resultante, a qual corresponde à média de todas as matrizes de confusão *Mclass*, calculadas para cada uma das diferentes partições.

4.4 Generalização do classificador

Seja qual for o classificador utilizado, ele precisa ser treinado usando exemplos de treinamento que permitam estimar os parâmetros mais adequados para a classificação. Como consequência, a performance do classificador depende tanto do número de elementos do conjunto de treinamento, como dos valores específicos dos exemplos, ou seja, da qualidade desses exemplos. Nesse sentido, a habilidade de generalização de um classificador se refere à sua capacidade em classificar padrões de teste que não tenham sido usados durante a etapa de treinamento.

Os problemas de generalização apresentam-se quando um classificador se especializa demais

em seus padrões de treinamento, ou quando utiliza mais informações (características) do que as necessárias. Uma baixa capacidade de generalização de um classificador pode ser atribuída a qualquer um dos seguintes fatores (Jain et al., 2000):

- O número de características é muito alto em relação ao número de elementos do conjunto de treino (*curse of dimensionality*);
- O número de parâmetros desconhecidos associados com o classificador é muito alto;
- O classificador é demasiadamente otimizado para o conjunto de treino específico, criando uma super-adequação dos parâmetros (*overtraining* ou *overfitting*).

Assim, o desempenho de um classificador depende: (1) da relação entre sua complexidade, (2) a qualidade do conjunto de treinamento (se é representativo dos dados) e (3) o número de características utilizadas. A Figura 4.2 mostra o fluxograma do processo de generalização do classificador, no qual se pode verificar que o núcleo está constituído pelo “banco de características” e pela “função-critério” (classificador). Esses dois componentes constituem-se na entrada aos dois processos paralelos: seleção de características e definição do tamanho mínimo do conjunto de treinamento, o qual determina o número mínimo de elementos com os quais o classificador produz resultados aceitáveis. Como é mostrado no fluxograma, ambos os processos trocam informações. Assim, as características selecionadas podem mudar de acordo com os dados de treinamento, os quais, por sua vez, dependem do subconjunto de características selecionadas. Após ter-se definido o subconjunto de características e o tamanho mínimo do conjunto de treinamento, prossegue-se com o processo de avaliação do desempenho do classificador. Esse processo tem por objetivo avaliar o comportamento de um classificador e, assim, poder compará-lo com outros classificadores.

4.4.1 O problema da dimensionalidade

O problema da dimensionalidade também é conhecido como “maldição da dimensionalidade” (*curse of dimensionality*), termo cunhado por Bellman (1961), que se refere ao rápido incremento no volume associado com a adição de dimensões extras a um espaço matemático. Esse problema é um obstáculo muito significativo na escolha da dimensionalidade ideal a ser adotada em um problema de reconhecimento de padrões. Ao contrário do senso comum, a probabilidade de classificação correta de uma regra de decisão não é diretamente proporcional ao aumento do número de características, mesmo que o número de exemplos de treinamento seja arbitrariamente grande e representativo da densidade subjacente.

Figura 4.2: Fluxograma do processo de generalização do classificador.

Sem dúvida, tem sido observado na prática que a adição de características pode até mesmo degradar o desempenho do classificador. De fato, um número maior de características implica na necessidade de um maior número de exemplos de treinamento (Valiant, 1984). Por outro lado, quando a dimensionalidade é alta, a simples redução de uma característica, sem critério, pode prejudicar a capacidade discriminativa do classificador devido à perda de informações discriminativas contidas nessa característica.

A curva da Figura 4.3 ilustra o problema da dimensionalidade. Pode-se identificar nessa curva três regiões no eixo da dimensionalidade com diferentes significados (Campos, 2001):

1. Na primeira região, compreendida entre 0 e m_1 , ocorre o comportamento mais esperado in-

Figura 4.3: Efeito da dimensionalidade na taxa de acerto de um classificador.

tuitivamente, pois a adição de características promove um aumento na taxa de acerto. Isso se deve ao fato de espaços com dimensões muito pequenas não possuírem informações suficientes para a distinção das classes de padrões. Assim, a adição de novas características, desde que sejam discriminativas, melhora os resultados da classificação.

2. A segunda região, compreendida entre m_1 e m_2 , é aquela em que é atingida uma estabilidade na taxa de acerto. Nessa região, a adição ou eliminação de características não altera substancialmente essa taxa. Para um problema de classificação, a melhor solução está na adoção da dimensionalidade m_1 , pois esse é o menor valor em que a taxa de acerto é máxima. A estabilização na taxa de acerto se deve ao fato de que as características importantes para se distinguir os padrões já foram incluídas na região anterior, e as características adicionais não são nem ruidosas e nem relevantes para a classificação.
3. A última região, posterior a m_2 , é a faixa em que de fato ocorre o problema da dimensionalidade. Note-se que o aumento no número de características, ao invés de melhorar a classificação, provoca redução na taxa de acerto. Isso pode ocorrer quando a adição de novas características inserir ruído e, assim, prejudicar a capacidade de discriminação.

Assim, para maximizar o desempenho do classificador, é necessário investigar qual é a dimensionalidade ideal para um determinado problema de reconhecimento de padrões. Para isso, pode ser aplicada uma estratégia simples de tentativa e erro em relação à dimensionalidade, usando um método de redução da dimensionalidade (incluindo extração e seleção de características) até que o ponto de máximo desempenho de um classificador seja atingido. Nessa estratégia, são realizados testes de redução de dimensionalidade para a obtenção de subespaços de características de vários tamanhos diferentes, até que seja obtida a dimensionalidade que minimiza o erro de classificação.

A pesar de não ser teoricamente clara a relação entre dimensionalidade e o tamanho do conjunto de treinamento, há outros fatores que, quando considerados, ofuscam a exatidão dessa relação, tais como a complexidade do classificador e o número de classes. Segundo Jain et al. (2000), é geralmente aceito usar, como número de exemplos de treinamento, pelo menos dez vezes o número de características. Finalmente, é fundamental se ter em mente que mais importante do que a dimensionalidade em si, é a escolha criteriosa das características mais discriminativas do objeto de estudo.

4.4.2 Seleção de características

O problema de seleção de características é definido (Jain et al., 2000) como uma técnica de otimização que, dado um conjunto de d características, tenta selecionar um subconjunto de m ($m < d$) características que maximiza uma função-critério de classificação, o que também representa uma forma de redução de dimensionalidade. A seleção de características é um problema do tipo *NP-hard* (Narendra e Fukunaga, 1977), o que significa que é de alta dificuldade de resolução.

Formalmente, dado um conjunto $\mathcal{Y}$ de d características, o algoritmo de seleção de características deve encontrar um subconjunto $\mathcal{X} \subseteq \mathcal{Y}$ tal que $|\mathcal{X}| = m$, em que $|\mathcal{X}|$ denota a cardinalidade de $\mathcal{X}$, e

$$J(\mathcal{X}) = \max_{\mathcal{X} \subseteq \mathcal{Y}, |\mathcal{X}|=m} J(\mathcal{X}), \quad (4.12)$$

em que $J(\cdot)$ é a função-critério. Uma maneira simples é definir $J(\mathcal{X}) = 1 - E$, sendo E a taxa ou probabilidade de erro de um classificador.

Dessa forma, um algoritmo eficiente de seleção de características poderá resultar na escolha de características altamente discriminativas implicando, assim, que um número relativamente pequeno de dimensões seja requerido para uma boa classificação. Em outras palavras, a seleção das melhores características pode levar a uma redução da dimensionalidade sem o comprometimento da capacidade discriminativa do classificador. Outra consequência positiva é que, ao se reduzir a dimensionalidade, também pode-se diminuir o tamanho do conjunto de treinamento.

Existem vários métodos de seleção de características. Jain e Zongker (1997) fizeram uma revisão das distintas abordagens e definiram uma classificação de métodos, a qual está mostrada na Figura 4.4. Kittler et al. (2001), por outro lado, definiram uma classificação baseada na disponibilidade e não disponibilidade de conhecimento a priori, onde os métodos de busca ótima ganham importância.

Neste trabalho foram usados métodos de busca ótima (destacados com uma linha descontínua

Figura 4.4: Esquema de classificação dos métodos de seleção de características. Adaptada de Jain e Zongker (1997).

na Figura 4.4), onde a família mais importante de métodos de busca exaustiva são os algoritmos *Branch-and-Bound* (Narendra e Fukunaga, 1977). A maior crítica aos métodos de busca ótima é sua alta complexidade computacional, o qual pode ser contornado acrescentando-se algumas técnicas inteligentes no processo de busca como, por exemplo, os métodos de busca sequencial para frente (SFS) e para trás (SBS) (Theodoridis e Koutroumbas, 1998; Campos, 2001), que são descritos a seguir:

Busca exaustiva

Este método também é conhecido como de força bruta, e consiste na avaliação de todas as possíveis combinações de características para encontrar o melhor subconjunto de n características, onde $n < d$, d é o número total de características. São verificadas todas as combinações possíveis para cada nível d de características, sendo que o número total de combinações para cada um dos subconjuntos S_i é expresso pela Equação 4.13.

$$S_i = \binom{d}{n} = \frac{d!}{(d-n)!n!} \quad (4.13)$$

Este método somente é aplicável para conjuntos com número pequeno de características, uma vez que o aumento da dimensionalidade acarreta um incremento exponencial dos recursos computacionais exigidos.

Busca seqüencial para frente

O método de busca seqüencial para frente (*Sequential Forward Selection* – SFS) é um método *bottom-up*. Dado um conjunto de características já selecionadas (inicialmente vazio), a cada geração é selecionada a característica que, unida ao subconjunto determinado na anterior iteração, produz o melhor resultado da função-critério. Essa característica é adicionada ao conjunto anterior de características, e uma nova iteração é então realizada.

A desvantagem desse método é que, uma vez que uma característica tenha sido adicionada, ela não pode ser descartada do subconjunto ótimo, o que pode provocar o chamado efeito *nesting*. O efeito *nesting* ocorre quando o subconjunto ótimo não contém elementos do conjunto já selecionado, o que impossibilita que seja obtido o conjunto de características ótimo. A principal vantagem da busca seqüencial para frente é o menor custo computacional quando se deseja obter conjuntos pequenos em relação ao total de características.

Busca seqüencial para trás

O algoritmo de busca seqüencial para trás (*Sequential Backward Selection* – SBS) é uma versão *top-down* do algoritmo anterior. A diferença entre SBS e SFS é que o SBS é iniciado com o conjunto de características completo (contendo todas as d características) e vai eliminando as menos importantes, ou seja, as que menos alteram a função-critério quando são eliminadas.

Assim como o método SFS, a desvantagem desse método é que, uma vez eliminada uma característica, ela não retornará ao subconjunto ótimo novamente. Como consequência, também pode ocorrer o efeito *nesting* caso o melhor subconjunto contenha uma ou mais características que tenham sido previamente eliminadas.

A principal vantagem desse método também é o menor o custo computacional, quando se deseja obter conjuntos grandes em relação ao total de características.

4.4.3 Função-critério

Conforme mencionado na introdução desta seção, uma das partes mais importantes na redução da dimensionalidade é a escolha de uma função-critério. Em seleção de características, o objetivo das funções-critério é minimizar o erro de classificação. Dessa forma, dado um conjunto de características ($\mathcal{X}$), um exemplo de funções-critério é: $J(\mathcal{X}) = 1 - E(\mathcal{X})$, sendo E a probabilidade de erro de um classificador usando $\mathcal{X}$ como conjunto de características. Essa probabilidade de erro pode ser determinada mediante a taxa de acerto de um classificador ou da distância entre as classes de padrões de treinamento no espaço de características.

Um critério amplamente utilizado é a taxa de acerto na classificação de um subconjunto de características. Basicamente, quando não se dispõe de informações a respeito da distribuição dos dados, a avaliação do classificador pode ser feita utilizando-se os padrões de treinamento e de teste no espaço determinado pelo conjunto de características. A taxa de acerto é utilizada como função-critério, de forma que, quanto maior a taxa de reconhecimento, melhor é o conjunto de características.

Neste trabalho, foram utilizados dois classificadores que fornecem como resultado uma taxa de acerto na classificação de um conjunto de elementos. Essas taxas de acerto foram usadas como critério de avaliação dos classificadores. Embora ambos os classificadores sejam baseados na teoria Bayesiana, o primeiro fornece resultados de similaridade (Seção 4.2.1), enquanto os resultados do segundo são probabilísticos (Seção 4.2.2).

4.4.4 Tamanho mínimo do conjunto de treinamento

Embora seja difícil se estabelecer teoricamente uma relação entre a probabilidade de erro, o número de exemplos de treinamento, o número de características e os parâmetros das densidades de classe condicional (como exposto na Seção 4.4.1) geralmente se admite que o número mínimo de exemplos de treinamento deve ser no mínimo dez vezes maior do que o número de características (Jain et al., 2000). Contudo, deve-se levar em consideração que a escolha de um classificador é um problema em si, e, assim, o tamanho mínimo do subconjunto de treinamento também depende do desempenho desse classificador (Jain et al., 2000). Além disso, fatores externos que influenciam a qualidade dos dados adquiridos podem resultar na necessidade de tamanhos maiores dos conjuntos de treinamento.

A abordagem experimental usada neste trabalho consistiu na extração aleatória de um conjunto de treinamento e um outro de teste, para cada classe, utilizando-se o método descrito no item 4.3. Diferentes proporções desses subconjuntos foram testadas, usando-se intervalos definidos por inte-

ros (ex. de 5 : 95 a 95 : 5). Além disso, para cada proporção treinamento:teste, foram feitas 100 amostragens aleatórias, reduzindo assim possíveis tendenciosidades na amostragem.

Cada amostra consistindo de subconjuntos de treinamento e teste, foi avaliada quanto à sua capacidade de treinar adequadamente o classificador, medida em termos de taxa de acerto. A média das taxas de acerto, obtidas para cada um dos pares de subconjuntos, foi considerada como a taxa global de classificação correta para aquela proporção de conjuntos de treinamento:teste. Esta abordagem foi recursivamente aplicada para todas as proporções treinamento:teste visando estimar qual delas resultava no melhor desempenho do classificador. A taxa de acerto da classificação foi utilizada como função-critério, gerando-se uma curva que indica a taxa de acerto para as distintas proporções de subconjuntos.

4.4.5 Avaliação do desempenho do classificador

No processo de reconhecimento de padrões, distintos classificadores podem ser utilizados no intuito de se obter uma melhor taxa de classificação correta. Uma forma rápida e simples de comparar o desempenho de classificadores distinto é através do cálculo da acurácia (Equação 4.14). Contudo, esse valor não necessariamente representa o comportamento do classificador, isto é, as condições sob as quais é usado o classificador, a distribuição das classes, e os custos de erro de classificação (Provost e Fawcett, 1997).

$$Acuracia = \frac{\text{n}^\circ \text{ de exemplos corretamente classificados}}{\text{n}^\circ \text{ total de exemplos positivos}} \quad (4.14)$$

Uma outra forma de se avaliar um classificador é através da geração das curvas ROC (*Receiver Operating Characteristics*)¹, uma técnica muito útil na organização de classificadores e visualização de desempenho. As curvas ROC têm sido usadas na tomada de decisões médicas (Zou, 2002; Swets, 1988; Swets et al., 2000) e, nos últimos anos, têm sido adotadas pelas comunidades de aprendizado de máquina e mineração de dados (Fawcett, 2006; Spackman, 1989; Bradley, 1997; Flach, 2004).

Muitos classificadores estão projetados para produzir só uma decisão binária como, por exemplo, **Y** ou **N** para cada instância/exemplo. Isso possibilita a existência de quatro possíveis saídas: exemplo positivo classificado como positivo (verdadeiro positivo), exemplo positivo classificado como negativo (falso negativo), exemplo negativo classificado como negativo (verdadeiro negativo) e, fi-

¹O nome ROC é pelo fato que os operadores de radares tinham que decidir se um ponto na tela representava um objetivo inimigo, uma nave amiga, ou simplesmente tratava-se de ruído. A teoria de detecção de sinal mensurava a habilidade dos operadores receptores a radar para fazer essas importantes distinções. Sua habilidade para realizar essa tarefa foi chamada de características operativas do receptor (*Receiver Operating Characteristics*).

nalmente, exemplo negativo classificado como positivo (falso positivo). Quando tal classificador discreto é aplicado a um conjunto de teste, produz uma única matriz de confusão a qual, por sua vez, corresponde a um ponto ROC. Dessa forma, um classificador discreto produz um simples ponto no espaço ROC.

Por outro lado, alguns classificadores como o Bayesiano ou uma rede neural, produzem uma probabilidade ou escore, um valor numérico que representa o grau no qual uma instância é membro de uma classe. Esses valores podem ser estritamente probabilísticos, em cujos casos aderem aos teoremas de probabilidade. Alternativamente, para escores não probabilísticos, um valor alto indica uma alta probabilidade.

Para estes classificadores, que fornecem um resultado hierarquizado (*ranking*), escore ou probabilidade, pode-se associar um valor de limiar (*threshold*), produzindo, assim, um classificador discreto (binário). Se a saída do classificador é maior do que o limiar, o classificador retorna um **Y**, e, se for menor, um **N**. Cada valor de limiar produz um ponto distinto no espaço ROC. Conceitualmente, o valor do limiar pode variar de $-\infty$ a $+\infty$, o que permite traçar uma curva no espaço ROC (Fawcett, 2006) e, portanto, a análise é feita independentemente da escolha desse valor. Quanto mais distante a curva estiver da diagonal principal, melhor será o desempenho do sistema de classificação para aquele domínio. Ao se comparar duas (ou mais) curvas, caso não haja nenhuma intersecção, a curva que mais se aproxima do ponto (0,1) é a de melhor desempenho. Caso haja intersecções, cada um dos sistemas tem uma faixa operacional na qual é melhor que o outro. Idealmente, a curva deveria ser convexa e sempre crescente.

Figura 4.5: Exemplos de curvas ROC.

A Figura 4.5 mostra dois exemplos de curvas ROC, comparando o desempenho de dois classificadores em cada uma. É importante notar na Figura 4.5(a) não há intersecções entre as curvas. Nesse caso, a curva superior representaria ao classificador de melhor desempenho. Já no caso da Figura 4.5(b), em que há uma intersecção entre as curvas próximo ao ponto (0.25, 0.60), a interpretação é de que o primeiro dos classificadores apresentou melhor desempenho até o ponto de intersecção, mas, a partir deste ponto, o segundo classificador apresentou melhor desempenho.

Curvas ROC multi-classe

Na análise de curvas ROC, assume-se que somente duas classes estão envolvidas no processo de classificação. De fato, a análise ROC é muito utilizada na tomada de decisões médicas nas quais se apresentam problemas de diagnóstico com duas classes – presença ou ausência de uma condição anormal. Os dois eixos representam a relação entre erros (falsos positivos) e benefícios (verdadeiros positivos) que ocorrem na discriminação de duas classes por um classificador. A análise é feita diretamente, devido à simetria existente em problemas com duas classes. Além disso, o desempenho do classificador é apresentado em duas dimensões, o que torna o resultado fácil de ser visualizado e interpretado.

Contudo, a presença de mais de duas classes é muito comum em problemas de reconhecimento de padrões. Isso faz com que a análise seja mais complexa se todas as dimensões forem analisadas juntas. Por exemplo, com c classes a matriz de confusão teria dimensão $c \times c$, contendo c classificações corretas (diagonal principal) e $n^2 - c$ possíveis erros (elementos não pertencentes à diagonal principal).

Uma abordagem para tratar casos com c classes consiste em se gerar c distintos grafos de curvas ROC, um para cada classe. Especificamente, se W é o conjunto de todas as classes, o grafo da curva ROC i identificaria o desempenho do classificador usando a classe ω_i como a classe positiva, e todas as classes restantes como a classe negativa, como exemplificado nas equações abaixo:

$$P_i = \omega_i \quad (4.15)$$

$$N_i = \bigcup_{j \neq i} \omega_j \in W \quad (4.16)$$

Finalmente, é importante ressaltar que as curvas ROC têm uma característica muito atraente, que é o fato de elas serem insensíveis ao desbalanceamento das classes (Fawcett, 2006).

Capítulo 5

Análise de distância a partir de dados morfológicos

5.1 Introdução

Neste capítulo será apresentada uma metodologia para o uso de dados morfológicos para uma análise de distância entre espécies de *Eimeria*. Tradicionalmente, a representação da morfologia dos organismos é realizada através de uma coleção arbitrária de medidas de distância, que envolvia o cálculo de ângulos, áreas, volumes, e outros dados quantitativos, extraídos especialmente de características anatômicas (*landmarks*). Esses métodos não levam em consideração a informação das relações espaciais entre as variáveis medidas.

Rohlf e Marcus (1993) proclamaram a denominada “revolução morfométrica” descrevendo uma abordagem que combina técnicas matemáticas e métodos estatísticos para a descrição morfológica, denominada de geometria morfométrica. Outros trabalhos na área abordam o mesmo problema como, por exemplo, sua aplicação em áreas da biologia e medicina (Bookstein, 1997; Marcus, 1996), análise estatística da forma (Dryden e Mardia, 1998), e propriedades dos espaços de forma (Small, 1996). A análise dos relacionamentos entre estatísticas das formas e espaços das formas foi reportada por (Rohlf, 1999). Com o incremento do poder computacional, outras técnicas também têm-se mostrado adequadas para a descrição e análise de formas, como os descritores de Fourier (Lestrel, 1997), curvatura multi-escala (Cesar Jr. e Costa, 1996), análise de dados morfométricos usando *wavelets* (Takemura et al., 2004), entre outros. Além desses métodos de representação de formas, é importante definir algumas medidas que permitam comparar quantitativamente duas formas, sendo uma das mais usadas as distâncias de Procrustes (Dryden e Mardia, 1998) – método

que aplica a sobreposição para comparar duas formas, definidas pela configuração de *landmarks*, o que também pode permitir descrever algumas diferenças entre várias configurações de landmarks *landmarks* (Rohlf e Slice, 1990).

Uma questão que tem recebido especial atenção, mas que ainda não foi resolvida, é como determinar o melhor modo de usar os dados morfométricos para uma reconstrução filogenética. Uma vez que os dados morfométricos fornecem uma rica descrição da forma morfológica, é natural pensar que esses dados poderiam ser usados para gerar árvores filogenéticas. No entanto, combinar essas duas disciplinas tem-se mostrado surpreendentemente difícil. Um dos maiores problemas está relacionado ao fato de que os dados morfométricos são contínuos e capturam o espaço de todas as possíveis variações de forma. Os métodos de inferência filogenética, em sua maioria desenvolvidos para seqüências biológicas, estão baseados em caracteres discretos que podem ser interpretados independentemente.

Uma segunda dificuldade diz respeito ao problema de como codificar os dados morfométricos em eventos evolutivos. Por exemplo, a simples discretização dos dados cria categorias artificiais e empíricas que não necessariamente guardam qualquer relação com eventos evolutivos. Por outro lado, o baixo número de caracteres morfológicos naturalmente discretos (ex. presença de algum orifício, barbatana, apêndice, etc.) limita seriamente as análises filogenéticas.

Outro problema pouco abordado na literatura é o uso de dados morfométricos para estabelecer relações filogenéticas de microorganismos, uma vez que a extração de dados morfométricos em um número razoável é ainda mais complexa do que em seres superiores. Por esta razão, dados moleculares se tornaram a alternativa mais comum para inferir relações filogenéticas entre microorganismos. Neste trabalho será relatado o uso de dados morfométricos de *Eimeria* spp. para análises de distância e sua comparação com análises filogenéticas realizadas com marcadores moleculares clássicos como o gene da subunidade 18S ribossômica e o citocromo *b* mitocondrial (Romano, 2004).

5.2 Inferência filogenética a partir de dados da forma

O uso de dados morfométricos na inferência filogenética tem sido motivo de várias discussões desde a década de 60, num debate “molecular” versus “morfologia” (Benton, 1999). Alguns autores afirmam que a morfologia não deveria mais ser utilizada na reconstrução filogenética (Scotland et al., 2003), o que foi rebatido por Jenner (2004) e Wiens (2004). Contudo, ambos os dados, moleculares e morfológicos, possuem propriedades únicas, e os dois tipos são válidos no esforço de se determinar o verdadeiro padrão filogenético (Patterson, 1987; Adoutte et al., 1998). Algumas

vantagens e desvantagens das duas abordagens são discutidas por Freeman e Herron (2004).

Existem várias metodologias para a reconstrução de árvores filogenéticas (Felsenstein, 2004). Os métodos mais usados são classificados em três grandes grupos (Nei e Kumar, 2000): métodos de distância, métodos de máxima parcimônia e métodos de máxima verossimilhança. Os métodos de distância utilizam matrizes de distância, enquanto são baseados em estados de caracteres. Swoford et al. (1996) faz uma outra distinção, entre métodos de algoritmos exatos e os de algoritmos heurísticos.

Um dos métodos mais populares da abordagem por distância é o denominado *unweighted pair-group*, que por sua vez, também é o mais simples para a reconstrução de árvores (Sokal e Michener, 1958; Nei, 1975). A técnica assume que as taxas de evolução são aproximadamente constantes entre as distintas linhagens evolutivas sob estudo. Essa suposição, contudo, nem sempre é verdadeira e, portanto, não funciona adequadamente em vários casos.

O método de máxima parcimônia usa dados de estado do caractere. O princípio da máxima parcimônia é procurar pela árvore que requer o menor número de mudanças evolutivas para explicar as diferenças entre as unidades taxonômicas sob estudo. Essa árvore é chamada de árvore de máxima parcimônia. Frequentemente mais de uma árvore com o mesmo número mínimo de mudanças pode ser encontrada e, dessa forma, nem sempre o método gera uma única árvore. Este método primeiramente foi desenvolvido para dados de sequência de aminoácidos (Eck e Dayhoff, 1966) e, posteriormente, foi modificado para ser usado em sequências de nucleotídeos (Fitch, 1977).

A abordagem de máxima verossimilhança está baseada em uma sólida teoria estatística. A primeira aplicação deste método na reconstrução de árvores foi feita por Cavalli-Sforza e Edwards (1967), que usou dados de frequências de genes. Posteriormente, Felsenstein (1973, 1981) desenvolveu algoritmos de máxima verossimilhança para dados de sequências de aminoácidos ou nucleotídeos. A máxima verossimilhança faz uso de um modelo probabilístico para o processo de substituição de nucleotídeos. Isso significa que é necessário especificar uma probabilidade de substituição de um nucleotídeo para outro em um intervalo de tempo por ramo. A verossimilhança de todos os sítios é o produto de todas as verossimilhanças de cada sítio individual, se todos os sítios nucleotídicos evoluem como caracteres independentes. Dado um conjunto de dados, calcula-se o valor de máxima verossimilhança para cada topologia de árvore, e esse procedimento é essencial para encontrar o comprimento dos ramos com os quais se consegue o maior valor para a função de verossimilhança. Finalmente, escolhe-se a topologia com o maior valor de verossimilhança como a melhor árvore, a qual é chamada de árvore de máxima verossimilhança.

Para o caso de dados morfológicos, a maior dificuldade está no fato de frequentemente serem variáveis contínuas, mas também outras questões ainda são discutidas (Wiens, 2000): (1) deveriam

esses caracteres ser considerados na análise filogenética, ou não?; (2) se forem considerados, que metodologias deveriam ser usadas na sua análise? As distintas técnicas usadas na filogenia por dados moleculares podem ser usadas para dados morfométricos mas, para isso, a informação deve ser qualitativa ou discreta (Wiens, 2000; Swiderski et al., 2002), o que significa que o problema de codificação de caracteres é fundamental nessa abordagem.

5.3 Discretização

A discretização é uma técnica para particionar atributos contínuos em um número de intervalos identificados com um valor discreto. Entre seus benefícios no processo de classificação podemos citar o incremento na velocidade de cálculo (Catlett, 1991) e o aumento na acurácia (Dougherty et al., 1995). Existem vários métodos de discretização, os quais podem ser classificados (Dougherty et al., 1995) como: globais vs. locais, supervisionados vs. não supervisionados, e estáticos vs. dinâmicos.

As duas abordagens mais simples correspondem à discretização por intervalos de mesmo tamanho e por intervalos de mesma frequência. Na primeira abordagem o intervalo é dividido em k pedaços de tamanho igual, enquanto na segunda os valores são divididos e ordenados em k intervalos, deixando cada intervalo com um número quase igual de elementos. No primeiro método existe o problema de perda de informação. Na segunda abordagem o objetivo é criar classes, separando os valores em conjuntos de tamanho parecido.

Uma outra abordagem de discretização usa a idéia da heurística de entropia mínima (Fayyad e Irani, 1993). Esse algoritmo define os intervalos das classes baseado no conceito de ganho de informação a cada corte possível. Os cortes possíveis são escolhidos a partir de um conjunto de dados de suporte.

5.4 Análise de distância usando dados morfológicos

O primeiro passo para construir as árvores de distância é gerar uma matriz de distância entre as distintas espécies. A Figura 5.1 mostra um esquema geral que permite calcular a matriz de distância, a qual considera como informações iniciais a matriz de médias e os intervalos de discretização de cada caractere.

Se o conjunto de dados possuir n espécies, algumas características poderão ser compartilhadas

Figura 5.1: Geração da matriz de distância a partir de dados morfométricos. Nesse exemplo, a discretização foi feita com intervalos de mesmo tamanho (a), gerando-se sete partições (b), aqui denominadas $P1$ a $P7$. As médias das medidas morfométricas de cada caractere (c), por espécie, foram designadas para as diferentes partições (b). A partir dessa classificação, foi gerada uma matriz de caracteres (d). A comparação de todas as combinações de pares de espécies foi usada para gerar uma matriz triangular de distância (e).

entre elas, enquanto outras serão distintas. Cada caractere quantitativo estará definido por uma faixa de valores. Assim, dentro de essa faixa, cada espécie terá uma distribuição de frequência da qual se pode extrair alguma informação como a média, moda, desvio padrão, entre outros. Na Figura 5.1(a) apresenta-se, como exemplo de procedimento, a distribuição de valores de um caractere sobre a qual foi praticado o processo de discretização por intervalos iguais $\{P1, \dots, P7\}$ (Figura 5.1(b)). Como são tomadas medidas de muitos indivíduos de cada espécie, utiliza-se valores das médias das medidas morfométricas criando-se assim, uma matriz de médias (Figura 5.1(c)).

A matriz de caracteres é calculada através da verificação do correspondente intervalo no qual está localizada a média. Por exemplo, o valor médio da espécie 1 ($Esp1$) no caractere $F1$ ($M(1,1)$) está localizado na partição 4 ($P4$). Assim, seu valor na matriz de caracteres corresponde a 4 (Figura 5.1(d)). O mesmo procedimento é seguido para cada um dos valores da matriz de médias. A dimensão da matriz de caracteres é, portanto, a mesma da matriz de médias. Finalmente, a matriz de distância (5.1(e)) contém os recíprocos do número de estados de caracteres compartilhado por quaisquer pares de espécies.

Capítulo 6

Resultados

6.1 Introdução

Neste trabalho foram conduzidos diversos experimentos orientados para a otimização do processo de diagnóstico automático da coccidiose aviária mediante a caracterização morfológica e análise de imagens digitais de oocistos dos parasitas. A partir disso propôs-se uma abordagem de diagnóstico dos microorganismos através de uma interface *web*, na qual imagens podem ser submetidas, pré-processadas e classificadas em tempo real. Essa abordagem integrada permitiu-nos estabelecer um sistema completo de diagnóstico parasitológico à distância que é, ao nosso conhecimento, o primeiro no mundo. Esse sistema, denominado COCCIMORPH, encontra-se disponível, no endereço *web* <http://puma.icb.usp.br/coccimorph/>. Nesse mesmo sítio de internet estão disponíveis instruções de uso, tutoriais, banco de imagens, *download* do programa, e demais materiais suplementares.

Além de estar focados no objetivo geral deste trabalho, os dados obtidos permitiram revisar a morfometria de parasitas do gênero *Eimeria* à luz de uma tecnologia mais moderna e não subjetiva. Além disso, o banco de imagens de *Eimeria* de galinha e de coelho foi colocado na internet e está publicamente disponível no endereço <http://puma.icb.usp.br/imagedb/>. Esse banco poderá ser extremamente útil no ensino de parasitologia, bem como constituir um repositório de imagens validadas que funcionaria como um “*golden standard*” para o teste de novos classificadores.

Finalmente, os dados morfométricos também permitiram iniciar estudos de análise de distância morfológica entre espécies de *Eimeria* e sua comparação com dados de filogenia molecular disponíveis na literatura ou produzidos em nosso laboratório. Neste capítulo são apresentados e analisados os resultados obtidos com as sete espécies de *Eimeria* da galinha doméstica, e também de onze

espécies de *Eimeria* que infectam o coelho doméstico. Essa abordagem pode ser facilmente estendida para espécies de *Eimeria* de outros animais hospedeiros ou de quaisquer microorganismos cuja morfologia é semelhante à de oocistos. Assim, outros protozoários com estágios de cistos, assim como ovos de helmintos, poderiam ser facilmente adaptados para o uso em nosso sistema.

6.2 Identificação de espécies de *Eimeria* de galinha

6.2.1 Conjunto de dados

O conjunto de dados está constituído de imagens das distintas espécies de *Eimeria* que infectam a galinha doméstica. Sempre que possível, foram usadas múltiplas cepas de cada espécie, coletadas em diferentes origens geográficas (Tabela 2.1). Isso foi feito para se reduzir o efeito de possíveis variações morfológicas entre cepas (variações intra-específicas). No total, o conjunto de dados usado nos distintos testes é composto por um total de 3891 imagens de oocistos das sete espécies. A Tabela 2.1 apresenta as espécies e cepas de *Eimeria* de galinha usadas neste trabalho. Os números apresentados correspondem àquelas imagens que foram pré-processadas, isto é, que foram segmentadas e tiveram suas características morfológicas extraídas com sucesso. Foi criado então um banco de imagens de *Eimeria* (*The Eimeria Image Database*) na internet, com acesso público, no endereço <http://puma.icb.usp.br/imagedb/>. Esse repositório compreende tanto as imagens passíveis de processamento quanto aquelas que foram refratárias. Essas últimas imagens não permitiram a obtenção de uma segmentação adequada por diferentes motivos, incluindo problemas de foco, contraste, iluminação, artefatos biológicos, etc. (ver item 2.3.2).

As distintas características morfológicas, extraídas a partir das imagens, compuseram um espaço de características 14-dimensional. A Tabela 6.1 apresenta as diferentes características com seus respectivos identificadores. Elas estão ordenadas de acordo com o identificador, os quais foram designados na medida em que cada nova característica era agregada ao espaço.

6.2.2 Classificadores

Neste trabalho, foram utilizadas duas abordagens de classificação como funções-critério. O primeiro classificador usa como função de verossimilhança a função normal (Seção 4.2.1). Nesse caso, o vetor de características de cada imagem foi avaliado por cada uma das funções discriminantes das respectivas espécies. O elemento foi designado na classe onde apresentou o mais alto valor de discriminação. Para o cálculo do grau de similaridade do elemento com a classe, calculou-se a

Tabela 6.1: Características morfológicas extraídas automaticamente das imagens de *Eimeria* spp. e seus respectivos identificadores.

ID	Característica
<i>F1</i>	Média da curvatura
<i>F2</i>	Desvio padrão da curvatura
<i>F3</i>	Entropia da curvatura
<i>F4</i>	Diâmetro maior (altura)
<i>F5</i>	Diâmetro menor (largura)
<i>F6</i>	Simetria em função do diâmetro maior
<i>F7</i>	Simetria em função do diâmetro menor
<i>F8</i>	Área
<i>F9</i>	Entropia da estrutura interna
<i>F10</i>	Segundo momento angular
<i>F11</i>	Contraste
<i>F12</i>	Momento da diferença inversa
<i>F13</i>	Entropia da matriz de co-ocorrência
<i>F14</i>	Excentricidade (altura/largura) ou <i>shape index</i>

distância Mahalanobis entre o elemento e o protótipo da espécie, o qual por sua vez foi calculado a partir do conjunto de treinamento (ver item 4.2.1). Esse classificador foi, portanto, denominado “classificador por similaridade” (Cs).

O segundo classificador (Seção 4.2.2), por outro lado, usa como função de verossimilhança a função de Dirichlet, e os resultados fornecem uma informação da probabilidade de um elemento pertencer a uma classe, daí este classificador ser denominado “classificador por probabilidade” (Cp).

6.2.3 Desempenho do conjunto de treinamento

Visando estimar o número mínimo de elementos requeridos para compor um conjunto de treino que resultasse numa alta taxa de classificação correta, foram conduzidos diversos experimentos. Considerando-se que o número de imagens por espécie não estava balanceado, foram extraídos aleatoriamente 320 elementos de cada classe, o que formou um novo conjunto de dados, constituído no total por 2240 exemplos de oocistos das sete espécies de *Eimeria*. Seguindo o processo de classificação descrito na Seção 4.3, para cada espécie, o conjunto de dados correspondente foi aleatoriamente dividido em dois grupos, o conjunto de treino e o conjunto de teste, em proporções relativas que variaram de 95% : 5% a 5% : 95%, respectivamente, usando-se intervalos definidos por números inteiros. Além disso, para cada proporção usou-se 100 partições aleatórias. A média da diagonal de cada uma das 100 matrizes de confusão resultantes foi considerada como a taxa de

acerto final (*score*) para cada uma das proporções de conjuntos treino:teste.

Figura 6.1: Efeito do tamanho do conjunto de treino na acurácia da classificação. Um total de 2240 imagens foram usadas na avaliação. O tamanho do conjunto de treino é representado pela porcentagem relativa ao total do conjunto de dados. O número absoluto de imagens também é apresentado (entre parênteses). Os resultados estão apresentados para o classificador por similaridade (linha cheia) e por probabilidade (linha tracejada).

Como pode ser observado na Figura 6.1, existe uma clara correlação entre o tamanho do conjunto de treino e a taxa de acerto (classificação correta). Os dois classificadores utilizados apresentam um comportamento similar. Para um conjunto de 2240 imagens, uma boa taxa de acerto é atingida com aproximadamente 30% das imagens do conjunto de dados. Considerando-se que o conjunto de dados está constituído por 2240 exemplos das distintas espécies de *Eimeria*, conclui-se que o número mínimo aceitável de elementos para o conjunto de treinamento seria de 96 imagens por cada espécie, perfazendo um total de 672 exemplos. De fato, esse resultado está em concordância com o conceito de que o número mínimo de exemplos de treinamento por classe deve ser dez vezes maior do que a dimensionalidade (o número de características), conforme exposto na Seção 4.4. Assim, em nosso caso, como empregamos um total de 14 características, o número de exemplos seria de 140. Contudo, como será apresentado adiante (ver itens 6.2.4 e 6.2.5), nossa abordagem já consegue com cerca de nove características uma taxa de acerto muito próxima da máxima obtida, o que seria compatível com o valor de 96 exemplos por classe.

6.2.4 Correlação de características

Para se obter um indicador que nos permitisse ter uma idéia do comportamento de cada uma das variáveis, foram calculadas as matrizes de correlação para cada espécie. Esses dados permitiram estabelecer o grau de correlação entre as características morfológicas permitindo, assim, remover características redundantes e, portanto, reduzir a dimensionalidade do espaço de características.

A Tabela 6.2 apresenta a matriz de correlação calculada para todas as espécies de *Eimeria*. A partir da análise desta matriz, observaram-se correlações entre algumas das características morfológicas como, por exemplo, a *F1* mantém uma alta correlação com *F3* ($-0,68\%$), *F4* ($-0,93\%$) e *F5* ($-0,83\%$). A *F13* mostra uma alta correlação com *F9* ($0,93\%$), *F10* ($-0,92\%$) e *F12* ($-0,90\%$).

Esses resultados mostram que se pode eliminar algumas das características sem perda significativa de informação, reduzindo-se assim o número de variáveis utilizadas. Observando-se os dados da Tabela 6.2, é possível evidenciar que as características *F1*, *F8*, *F9*, *F12* e *F13* poderiam ser descartadas, limitando, assim, o número de variáveis originais para nove características. Essa informação, por ser rápida e fácil de calcular, serve de base para estabelecer o conjunto mínimo de características que permitem uma discriminação eficiente das espécies.

6.2.5 Seleção de características

A abordagem descrita no item anterior (Seção 6.2.4) pode fornecer uma primeira indicação de que características são redundantes e, portanto, poderiam em princípio ser descartadas. Contudo, este tipo de análise de grau de correlação entre características não leva em consideração os classificadores. Assim, decidiu-se também realizar um teste ótimo de seleção de características, visando se encontrar a sua melhor combinação, e utilizar cada classificador como função-critério. Para isso, foi aplicada a abordagem SFS (Seção 4.4.2) para se calcular a melhor combinação de características. Para cada combinação gerada pelo algoritmo SFS, foi montado um subconjunto de dados respectivo, o qual foi usado como entrada para o processo de classificação (descrito na Seção 4.3). Para se determinar a média geral da taxa de classificação correta, cada subconjunto foi aleatoriamente dividido em um conjunto de treino (30%) e um conjunto de teste (70%). Esse procedimento, por sua vez, foi repetido 100 vezes (procedimento de retroalimentação da Figura 4.1). Para cada dimensão, foi determinada e selecionada a combinação que resultava na maior taxa de acerto (de acordo com o algoritmo SFS), variando de dois até catorze características combinadas.

Tabela 6.2: Média das matrizes de correlação das sete espécies de *Eimeria* de galinha

Nível	<i>F1</i>	<i>F2</i>	<i>F3</i>	<i>F4</i>	<i>F5</i>	<i>F6</i>	<i>F7</i>	<i>F8</i>	<i>F9</i>	<i>F10</i>	<i>F11</i>	<i>F12</i>	<i>F13</i>	<i>F14</i>
<i>F1</i>	1,00													
<i>F2</i>	-0,09	1,00												
<i>F3</i>	-0,68	0,16	1,00											
<i>F4</i>	-0,93	0,16	0,59	1,00										
<i>F5</i>	-0,83	-0,17	0,65	0,65	1,00									
<i>F6</i>	0,01	0,32	0,09	0,04	-0,10	1,00								
<i>F7</i>	-0,03	0,32	-0,03	0,15	-0,15	0,23	1,00							
<i>F8</i>	-0,98	0,01	0,68	0,92	0,88	0,02	0,00	1,00						
<i>F9</i>	0,00	0,06	-0,05	0,01	-0,03	0,01	0,00	-0,01	1,00					
<i>F10</i>	-0,07	-0,12	0,02	0,04	0,09	-0,06	-0,02	0,08	-0,87	1,00				
<i>F11</i>	0,32	0,15	0,00	-0,31	-0,28	0,14	0,08	-0,32	0,24	-0,40	1,00			
<i>F12</i>	-0,12	-0,10	0,05	0,09	0,13	-0,05	-0,03	0,12	-0,73	0,89	-0,50	1,00		
<i>F13</i>	0,00	0,06	-0,02	0,02	0,02	0,01	0,00	0,00	0,93	-0,92	0,40	-0,90	1,00	
<i>F14</i>	-0,18	0,39	-0,03	0,46	-0,34	0,16	0,38	0,11	0,06	-0,06	-0,04	-0,04	0,06	1,00

Seleção de características usando o classificador por similaridade como função-critério

A Tabela 6.3 apresenta os resultados do experimento feito com 3891 elementos. Como pode ser observado, a melhor combinação de duas características ($F4$ e $F5$) resultou em uma taxa de acerto de 76,45%. O mais alto valor de classificação correta (85,39%) foi obtido com a combinação de um total de treze características.

Tabela 6.3: Seleção de características usando SFS e o classificador por similaridade para 3891 elementos

Nível	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	$F9$	$F10$	$F11$	$F12$	$F13$	$F14$	Acerto
2				×	×										76,45
3				×	×		×								79,13
4				×	×		×				×				80,15
5				×	×		×		×		×				81,47
6				×	×		×		×		×		×		82,92
7				×	×		×		×		×	×	×		84,37
8			×	×	×		×		×		×	×	×		85,08
9			×	×	×		×	×	×		×	×	×		85,00
10			×	×	×		×	×	×	×	×	×	×		85,09
11	×		×	×	×		×	×	×	×	×	×	×		85,14
12	×	×	×	×	×		×	×	×	×	×	×	×		85,29
13	×	×	×	×	×		×	×	×	×	×	×	×	×	85,39
14	×	×	×	×	×	×	×	×	×	×	×	×	×	×	85,18

Comparando-se os dados da Tabela 6.3 com os da Tabela 6.2, não se observa em princípio uma grande concordância. De fato, ambas as combinações, de $F5$ ou $F6$ características (Tabela 6.3), excluem as características $F1$, $F2$, $F3$, $F6$, $F8$, $F10$, $F12$ e $F14$, enquanto no teste de correlação concluiu-se que poderiam ser descartadas as características $F1$, $F8$, $F9$, $F12$ e $F13$. Essa suposta discrepância pode ser explicada para os diferentes casos. Por exemplo, as características $F2$ e $F6$ e $F14$, embora não apresentem correlação significativa com outras características (Tabela 6.2), não parecem ser importantes em termos discriminativos (Tabela 6.3) e, por isso, poderiam ser descartadas pelo segundo teste. A característica $F9$, por sua vez, apresenta correlação alta com a $F10$, $F12$ e $F13$, sendo por isso sugerido o seu descarte de acordo com o teste de correlação. Contudo, a análise por SFS, demonstrou que o uso da característica $F9$ e o descarte da $F10$ resulta em uma taxa de acerto maior, o que leva a concluir que características com baixa correlação mútua são desejáveis, contudo, podem ser inúteis se não forem discriminativas.

Seleção de características usando o classificador por probabilidade como função-critério

A Tabela 6.4 apresenta os resultados da seleção de características usando-se o algoritmo SFS, com o classificador por probabilidade como função-critério. A combinação das duas características mais discriminantes ($F5$ e $F7$) resultou em uma taxa de acerto de 69,51%. O maior valor de classificação correta (81,46%) foi obtido com a combinação de onze características.

Tabela 6.4: Seleção de características usando SFS e o classificador por probabilidade para 3891 elementos

Nível	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	$F9$	$F10$	$F11$	$F12$	$F13$	$F14$	Acerto
2					×		×								69,51
3				×	×		×								74,80
4				×	×		×							×	77,69
5				×	×		×				×			×	79,61
6				×	×		×			×	×			×	81,02
7				×	×		×		×	×	×			×	81,05
8				×	×	×	×		×	×	×			×	81,20
9			×	×	×	×	×		×	×	×			×	81,15
10		×	×	×	×	×	×		×	×	×			×	81,28
11		×	×	×	×	×	×		×	×	×		×	×	81,46
12		×	×	×	×	×	×		×	×	×	×	×	×	80,88
13		×	×	×	×	×	×	×	×	×	×	×	×	×	80,97
14	×	×	×	×	×	×	×	×	×	×	×	×	×	×	80,10

Comparando-se os dados da Tabela 6.4 com os da Tabela 6.2, ao contrario do que aconteceu com os dados da Tabela 6.3, observa-se uma maior concordância. De fato, com a combinação de 6 características, consegue-se eliminar aquelas também sugeridas através da matriz de correlação (características $F1$, $F8$, $F9$, $F12$ e $F13$).

Na Tabela 6.4 ganha destaque a característica $F14$, que mostra-se discriminativa, o que não acontece na Tabela 6.3. Em relação às características $F4$, $F5$, $F7$ e $F11$, elas se mantêm como as mais discriminativas nas duas Tabelas 6.3 e 6.4. Estas características estão referidas especialmente à representação do tamanho dos oocistos ($F4$ e $F5$), sua simetria no eixo menor ($F7$) e o contraste da estrutura interna ($F11$).

6.2.6 Discriminação de espécies

Depois de ter sido feita uma seleção de características (Seção 6.2.5) e definido o tamanho do conjunto de treino (Seção 6.2.3), procedeu-se a uma análise da discriminação de espécies. Conforme pode ser observado na Figura 6.2, foram realizadas três baterias de testes com cada classificador.

Na primeira, todos os elementos foram utilizados para treinar e testar os classificadores por similaridade (Cs-1) e por probabilidade (Cp-1), ou seja, os conjuntos de treinamento e de teste eram os mesmos. Na segunda rodada de testes, foram utilizados 30% dos elementos disponíveis para treinar o classificador e o restante como amostras-teste (Cs-2 e Cp-2). A extração desse subconjunto de treinamento (30%) foi feita de forma aleatória, com 100 replicatas, conservando-se a proporção de 30% treinamento para 70% teste, para cada espécie, resultando numa matriz média de classificação final (procedimento explicado na Seção 4.3).

Figura 6.2: Comparação das taxas de acerto dos classificadores por similaridade (Cs) e probabilístico (Cp) frente ao conjunto de 3891 imagens das sete espécies de *Eimeria* de galinha doméstica. A avaliação foi feita utilizando-se os métodos de “Todos contra Todos” (Cs-1 e Cp-1), “Partição aleatória” de 30% (Cs-2 e Cp-2) e *Leave One Out* (Cs-3 e Cp-3).

Finalmente, a terceira bateria de testes consistiu na aplicação da estratégia *leave-one-out* (Cs-3 e Cp-3). Nessa estratégia, para cada classe, o conjunto de treinamento foi inicialmente composto por todos os elementos, menos o primeiro, o qual foi utilizado para testar a classificação. Na segunda iteração de testes, o conjunto de treinamento foi composto por todos os padrões menos o segundo, o qual foi utilizado para teste. Esse processo foi repetido até que todos os indivíduos de cada classe tivessem sido utilizados para testar o classificador, usando-se sempre o restante das amostras para o treinamento. No final, a taxa média de acerto foi calculada.

A Figura 6.2 apresenta as médias de classificação obtidas com as distintas estratégias, e com os dois classificadores (Cs e Cp). Verifica-se que o classificador por similaridade (Cs) apresenta as melhores taxas de classificação correta, com uma média ao redor de 85%, enquanto o classificador por probabilidade atinge uma taxa de cerca de 80%. A comparação dos diferentes métodos de validação dos classificadores mostrou diferenças muito pequenas. Por outro lado, com qualquer dos

métodos de validação, bem como do classificador, observou-se um desvio-padrão relativamente alto, variando de 75-95% no classificador por similaridade e 65-95% no por probabilidade. Estes altos valores indicam uma alta variabilidade de acerto dependendo dos conjuntos de treinamento/teste empregados, sugerindo que existem grandes variações morfológicas entre os indivíduos de cada classe.

Uma apresentação mais detalhada dos resultados dos classificadores pode ser visualizada através de matrizes de confusão, nas quais é possível se observar as taxas de acerto para cada espécie, assim como as classificações erradas. As tabelas 6.5, 6.6 e 6.7 apresentam matrizes de confusão para as três estratégias de validação, usando-se o classificador por similaridade (Cs). As tabelas 6.8, 6.9 e 6.10 apresentam as respectivas matrizes de confusão usando-se o classificador por probabilidade (Cp).

Conforme pode ser visto nas tabelas 6.5, 6.6 e 6.7, não houve diferenças muito significativas entre os três métodos de validação testados com o classificador por similaridade. A média global (média da diagonal das matrizes) das taxas de classificação correta foi de cerca de 85%. *E. maxima* e *E. mitis* foram as espécies que apresentaram as melhores taxas de acerto. Uma das razões que explicam este resultado é o fato de que essas espécies apresentam os valores de diâmetro maior mais discrepantes em relação ao restante das espécies de *Eimeria*, sendo *E. maxima* a de maior diâmetro (daí o nome da espécie) e *E. mitis* a de menor. De fato, a Tabela 6.3, de seleção de características, mostra nitidamente que as características *F4* e *F5* (diâmetros maior e menor respectivamente) são as mais discriminativas, referendando essa hipótese.

Na análise da discriminação com o classificador por probabilidade (Tabelas 6.8, 6.9 e 6.10), também não se observaram grandes diferenças entre os métodos de validação. A taxa de acerto global foi de cerca de 80%, ligeiramente inferior à obtida pelo classificador por similaridade. Essa taxa inferior foi particularmente mais notada naquelas espécies com maior dificuldade de diferenciação como *E. praecox*, *E. tenella* e *E. necatrix*.

Tabela 6.5: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por similaridade e validação pelo método “Todos contra Todos” (Cs-1). Média da diagonal = 86,26%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	81,76	0,00	0,00	1,57	0,00	1,57	15,09
<i>E. maxima</i>	321	0,00	99,38	0,62	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,00	0,24	95,93	0,00	0,72	2,63	0,48
<i>E. mitis</i>	757	0,66	0,00	0,00	92,73	2,51	0,26	3,83
<i>E. praecox</i>	747	0,00	0,00	3,88	6,16	73,09	5,09	11,78
<i>E. tenella</i>	608	0,49	0,00	2,47	0,16	3,95	80,76	12,17
<i>E. necatrix</i>	404	2,48	0,00	0,50	3,71	3,96	9,16	80,20

Tabela 6.6: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por similaridade e validação por “partição aleatória” de 30% (Cs-2). Média da diagonal = 84,87%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	86,93	0,01	0,00	1,04	0,18	1,90	9,95
<i>E. maxima</i>	321	0,00	99,05	0,95	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,00	0,69	93,30	0,00	1,25	4,28	0,48
<i>E. mitis</i>	757	1,29	0,00	0,00	91,97	2,97	0,49	3,28
<i>E. praecox</i>	747	0,02	0,02	2,90	6,47	75,15	6,53	8,90
<i>E. tenella</i>	608	0,81	0,00	2,19	0,48	5,47	80,42	10,64
<i>E. necatrix</i>	404	7,19	0,00	0,48	5,23	5,79	14,04	67,28

Tabela 6.7: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por similaridade e validação pelo método “leave-one-out” (Cs-3). Média da diagonal = 85,49%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	81,76	0,00	0,00	1,73	0,16	1,73	14,62
<i>E. maxima</i>	321	0,00	99,07	0,94	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,00	0,48	94,02	0,00	0,96	4,07	0,48
<i>E. mitis</i>	757	0,66	0,00	0,00	92,74	3,04	0,26	3,30
<i>E. praecox</i>	747	0,00	0,00	3,08	6,69	75,77	5,62	8,84
<i>E. tenella</i>	608	0,49	0,00	2,47	0,33	4,61	80,59	11,51
<i>E. necatrix</i>	404	2,97	0,00	0,50	4,70	5,20	12,13	74,51

Tabela 6.8: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método “Todos contra Todos” (Cp-1). Média da diagonal = 80,78%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	85,22	0,00	0,00	2,20	0,16	2,20	10,22
<i>E. maxima</i>	321	0,00	99,07	0,93	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,00	0,96	89,71	0,00	3,11	6,22	0,00
<i>E. mitis</i>	757	1,06	0,00	0,00	91,68	2,11	0,13	5,02
<i>E. praecox</i>	747	0,13	0,00	5,09	5,76	66,00	8,84	14,19
<i>E. tenella</i>	608	0,66	0,00	2,96	0,16	5,10	70,89	20,23
<i>E. necatrix</i>	404	6,68	0,00	0,99	5,69	6,19	17,57	62,87

Tabela 6.9: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método de “partição aleatória” de 30% (Cp-2). Média da diagonal = 79,26%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	84,57	0,00	0,04	2,18	0,09	2,70	10,42
<i>E. maxima</i>	321	1,37	96,19	2,44	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,27	1,19	88,59	0,00	3,90	5,96	0,09
<i>E. mitis</i>	757	1,08	0,00	0,00	90,76	2,60	0,16	5,40
<i>E. praecox</i>	747	0,16	0,01	4,67	5,78	65,44	9,06	14,87
<i>E. tenella</i>	608	0,88	0,00	2,75	0,29	5,38	69,99	20,72
<i>E. necatrix</i>	404	7,41	0,00	1,01	5,67	6,19	20,46	59,25

Tabela 6.10: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da galinha utilizando-se um classificador por probabilidade e validação pelo método “leave-one-out” (Cp-3). Média da diagonal = 80,10%.

Espécies	Número oocistos	Espécies de oocistos						
		<i>E.ace</i>	<i>E.max</i>	<i>E.bru</i>	<i>E.mit</i>	<i>E.pra</i>	<i>E.ten</i>	<i>E.nec</i>
<i>E. acervulina</i>	636	84,43	0,00	0,16	2,20	0,16	2,20	10,85
<i>E. maxima</i>	321	0,00	98,44	1,56	0,00	0,00	0,00	0,00
<i>E. brunetti</i>	418	0,00	0,96	89,24	0,00	3,11	6,70	0,00
<i>E. mitis</i>	757	1,06	0,00	0,00	91,41	2,25	0,13	5,15
<i>E. praecox</i>	747	0,13	0,00	5,09	5,89	65,33	8,97	14,59
<i>E. tenella</i>	608	0,66	0,00	2,96	0,16	5,10	70,23	20,89
<i>E. necatrix</i>	404	7,18	0,00	0,99	5,69	6,19	18,32	61,63

6.2.7 Análise comparativa do desempenho dos classificadores

Para a comparação dos dois classificadores, foram utilizadas 3891 imagens cuja identidade havia sido determinada por métodos biológicos. A íntegra da classificação comparativa de cada um dos elementos do banco de dados, com os dois classificadores usados, pode ser visualizada no endereço <http://puma.icb.usp.br/coccimorph/classification/>.

A Figura 6.3 ilustra os resultados da classificação de uma parte dos elementos das espécies *E. praecox* e *E. necatrix*, onde cada imagem está representada por um identificador (coluna *indexDB*), seguida dos resultados de classificação para cada uma das sete espécies (colunas *1.ACE*, *2.MAX*, *3.BRU*, *4.MIT*, *5.PRA*, *6.TEN* e *7.NEC*). A coluna *Class* apresenta o número correspondente à classe de maior similaridade (ou probabilidade no caso do classificador por probabilidade), o que corresponde, portanto, ao diagnóstico do classificador. A coluna *Orig*, por outro lado, apresenta o número correspondente à espécie original da amostra (determinada por métodos biológicos). A mesma série de colunas é apresentada na parte direita da figura, correspondendo aos resultados relativos ao classificador por probabilidade.

Para exemplificar, a imagem *PRA102* (a segunda da lista), usando-se o classificador Cs, foi diagnosticada como sendo da espécie 7 (*E. necatrix*) com uma taxa de similaridade de 47,21%, mas se tratava de um indivíduo da classe 5 (*E. praecox*). De fato, a taxa de similaridade para *E. praecox* foi muito próxima, com 45,48%. Usando-se o classificador Cp, esse mesmo elemento foi corretamente classificado como pertencente à classe 5, com uma probabilidade de 51,22%. Contudo, a segunda maior probabilidade foi para *E. necatrix*, com 46,46%. Trata-se, portanto, de um típico exemplo de uma imagem cuja morfologia poderia ser considerada limítrofe (*borderline*) entre *E. praecox* e *E. necatrix*.

Um exemplo de classificação incorreta em ambos os classificadores é o da imagem *NEC102* (segunda imagem do segundo grupo da figura). A imagem é originalmente da classe 7 (*E. necatrix*), porém, o classificador Cs a designou como espécie 5 (*E. praecox*) com taxa de 57,96% e o classificador Cp como espécie 4 (*E. mitis*), com uma taxa de 99,12%.

No intuito de se avaliar o desempenho e comportamento dos classificadores ao longo do espaço de dados, foi gerado um conjunto de curvas ROC multi-classe (Figura 6.4), produzidas com os elementos verdadeiros positivos e falsos positivos (veja seção 4.4.5).

Uma vez que as curvas ROC são geradas a partir da classificação dos elementos considerando-se duas classes, o problema de classificação das sete espécies de *Eimeria* de galinha foi analisado através da geração de um grafo para cada espécie. Assim, o grafo da curva ROC de *E. acervulina* identificou o desempenho do classificador usando *E. acervulina* como a classe positiva e todas as

Classification with Similarity Approach												
FileName	IndexDB	1.ACE	2.MAX	3.BRU	4.MIT	5.PRA	6.TEN	7.NEC	Class	Orig	IndexDB	1.ACE
<u>PRA101</u>	1231	0	0	0	70.17	84.02	60.08	63.46	5	5	1231	0
<u>PRA102</u>	1232	0	0	0	22.25	46.48	24.04	47.21	7	5	1232	0
<u>PRA103</u>	1233	0	0	0	64.94	73.82	71.96	62.27	5	5	1233	2.28E-005
<u>PRA106</u>	1234	0	0	0	61.3	62.36	67.93	68.66	5	5	1234	2.34E-005
<u>PRA106</u>	1235	0	0	0	68.36	79.84	39.81	59.36	5	5	1235	8.74E-005
<u>PRA107</u>	1236	0	0	0	80.81	84.66	58.78	66.82	4	5	1236	0
<u>PRA108</u>	1237	0	0	0	73.76	86.61	60.25	68.37	5	5	1237	0
<u>PRA109</u>	1238	0	0	0	65.49	82.96	60.03	64.97	5	5	1238	0
<u>PRA111</u>	1239	26.23	0	0	66.05	87.16	68.8	81.6	5	5	1239	0
<u>PRA113</u>	1240	0	0	0	52.66	74.86	56.33	71.42	5	5	1240	0
<u>PRA116</u>	1241	0	0	0	83.51	71.78	0	52.41	4	5	1241	0
<u>PRA117</u>	1242	0	0	0	72.5	84.77	44.65	65.16	5	5	1242	0
<u>PRA118</u>	1243	0	0	0	80.93	39.27	50.51	5	5	5	1243	0
<u>PRA119</u>	1244	0	0	0	66.29	0	0	0	5	5	1244	0
<u>PRA120</u>	1245	69.2	0	0	43.07	66.74	67.95	81.19	7	5	1245	0.41
<u>PRA121</u>	1246	0	0	36.45	0	68.45	0	0	5	5	1246	0
<u>PRA125</u>	1247	0	0	0	67.86	84.97	60.31	63.92	5	5	1247	0
Classification with Probabilistic Method												
FileName	IndexDB	1.ACE	2.MAX	3.BRU	4.MIT	5.PRA	6.TEN	7.NEC	Class	Orig	IndexDB	1.ACE
<u>NEC101</u>	1919	65.18	0	0	37.12	0	0	71.03	7	7	1919	18.18
<u>NEC102</u>	1920	0	0	0	17.3	57.96	0	36.9	5	7	1920	0
<u>NEC103</u>	1921	65.43	0	0	73.86	75.01	56.17	80.61	7	7	1921	40.73
<u>NEC106</u>	1922	26.29	0	0	71.91	61.29	0	51.13	4	7	1922	0.01
<u>NEC106</u>	1923	68.9	0	0	89.31	49.76	0	52.58	4	7	1923	0
<u>NEC107</u>	1924	5.02	0	0	76.82	83.88	74.29	75.95	5	7	1924	0
<u>NEC108</u>	1925	65.74	0	0	0	0	0	41.77	1	7	1925	77.24
<u>NEC109</u>	1926	19.64	0	0	0	67.68	63.18	74.15	7	7	1926	1.82E-005
<u>NEC110</u>	1927	0	0	7.45	0	84.37	75.04	67.3	5	7	1927	0
<u>NEC111</u>	1928	24.12	0	0	0	56.61	61.49	89.56	7	7	1928	1.97E-005
<u>NEC112</u>	1929	46.94	0	0	0	0	0	77.74	7	7	1929	0.04
<u>NEC113</u>	1930	47.42	0	0	40.58	0	0	68.3	7	7	1930	98.22
<u>NEC114</u>	1931	68.72	0	0	68.5	0	0	60.09	4	7	1931	93.53
<u>NEC116</u>	1932	0	0	0	33.21	67.51	81.49	78.29	6	7	1932	1.18
<u>NEC116</u>	1933	0	0	0	32.75	84.52	78.76	85.87	7	7	1933	5.45E-008
<u>NEC117</u>	1934	0	0	0	29.32	88.9	47.98	56.86	5	7	1934	0
<u>NEC118</u>	1935	0	0	0	47.34	83.21	83.14	86.46	7	7	1935	0

Figura 6.3: Comparação dos resultados de classificação para alguns elementos de *E. praecox* e *E. necatrix* (<http://puma.icb.usp.br/coccinorph/classification/>).

(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figura 6.4: Avaliação comparativa dos classificadores por similaridade (linha azul) e por probabilidade (linha vermelha) através de curvas ROC. A linha preta indica a diagonal de referência. Espécies: (a) *E. acervulina*, (b) *E. maxima*, (c) *E. brunetti*, (d) *E. mitis*, (e) *E. praecox*, (f) *E. tenella*, e (g) *E. necatrix*.

espécies restantes como a classe negativa. A estratégia *leave-one-out* foi aplicada para classificar todos os elementos do conjunto de dados, com cada um dos dois classificadores.

A partir dos resultados de classificação de cada elemento do banco de dados com os dois classificadores, procedeu-se com a geração das curvas ROC para cada espécie. Para isso, os elementos “verdadeiros positivos” e “falsos positivos” de cada espécie foram selecionados e usados na geração das curvas ROC (Figura 6.4).

Por exemplo, para se gerar a curva ROC da espécie *E. acervulina*, usando-se o classificador Cs, foram selecionados como verdadeiros positivos todos aqueles elementos que foram corretamente classificados como *E. acervulina* ($Class = 1$ e $Orig = 1$). Posteriormente, foram selecionados todos os elementos falsos positivos, que correspondiam àqueles elementos classificados como *E. acervulina*, mas cuja classe original pertencia a uma outra espécie ($Class = 1$ e $Orig \neq 1$). Cada um desses elementos selecionados foi extraído junto com suas respectivas taxa de classificação. A seguir, os elementos selecionados (verdadeiros positivos e falsos positivos) foram ordenados de forma decrescente em função da taxa de classificação, o que constituiu a entrada para o traçado da curva ROC. O mesmo procedimento foi aplicado com os resultados obtidos com o classificador Cp. As curvas resultantes de cada classificador, estão mostradas na Figura 6.4(a), onde a curva azul corresponde ao classificador Cs e a vermelha ao classificador Cp. Da mesma forma, foram geradas as curvas ROC para cada uma das demais espécies, também mostradas na Figura 6.4. Quanto maior a área sob a curva ROC, melhor é o desempenho do classificador para aquela espécie.

Analisando-se a Figura 6.4, verifica-se nas curvas ROC que o classificador por probabilidade (Cp) apresentou um melhor desempenho do que o classificador por similaridade (Cs), o que representa um resultado muito interessante, considerando-se que em todas as matrizes de classificação o Cs sempre apresentou melhores taxas de acerto do que Cp. A curva ROC nos mostra que no Cp há casos de falsos positivos com altas taxas de probabilidades, porém são casos raros. Na maioria das vezes em que o Cp atribui uma alta probabilidade para alguma classe, trata-se de um verdadeiro positivo e, portanto, essa classificação é muito confiável. Por outro lado, no caso do Cs, a taxa global de acerto é ligeiramente maior. Contudo, muito freqüentemente, o valor da segunda classe mais similar é muito próximo ao valor da primeira. A taxa de verdadeiros positivos nesses casos tende a ser relativamente mais baixa e, portanto, menos confiável.

6.3 Identificação de espécies de *Eimeria* de coelho

Embora o trabalho todo tenha sido padronizado com as espécies de *Eimeria* de galinha, decidimos estendê-lo para as onze espécies de *Eimeria* que infectam o coelho. Estas amostras de *Eimeria* nos foram gentilmente fornecidas pelo Dr. Michal Pakandl (Academia de Ciências da República Tcheca). A Figura 6.5 apresenta uma micrografia dos oocistos das distintas espécies de *Eimeria* de coelho. Verifica-se que algumas espécies possuem uma morfologia muito distinta daquela observada nas espécies que infectam a galinha.

Figura 6.5: Micrografia de oocistos das onze espécies de *Eimeria* de coelho. Espécies: (a) *E. exigua*, (b) *E. perforans*, (c) *E. piriformis*, (d) *E. flavescens*, (e) *E. irresidua*, (f) *E. stiedai*, (g) *E. intestinalis*, (h) *E. media*, (i) *E. vej dovskyi*, (j) *E. coecicola* e (k) *E. magna*.

Aplicaram-se procedimentos de aquisição, pré-processamento e classificação de imagens (os mesmos aplicados à *Eimeria* de galinha) no intuito de testar a abordagem proposta nesta tese e verificar o comportamento com o novo conjunto de imagens. A Tabela 6.11 apresenta as distintas espécies, sua procedência e o número de imagens adquiridas, totalizando 2167 elementos.

Tabela 6.11: Origem geográfica das cepas de *Eimeria* de coelho e respectivo número de imagens utilizadas neste trabalho.

Nome	Origem	Número
<i>E. coecicola</i>	Nouzilly, França	191
<i>E. exigua</i>	Nouzilly, França	282
<i>E. flavescens</i>	Ceske Budejovice, República Tcheca	186
<i>E. intestinalis</i>	Nouzilly, França	127
<i>E. irresidua</i>	Ceske Budejovice, República Tcheca	209
<i>E. magna</i>	Nouzilly, França	291
<i>E. media</i>	Nouzilly, França	199
<i>E. perforans</i>	Nouzilly, França	110
<i>E. piriformis</i>	Ceske Budejovice, República Tcheca	133
<i>E. stiedai</i>	Ceske Budejovice, República Tcheca	156
<i>E. vejdoskyi</i>	Ceske Budejovice, República Tcheca	283

6.3.1 Tamanho do conjunto de treino e seleção de características

Por se tratar de um maior número de espécies, verificamos se o número mínimo de imagens necessárias para o conjunto de treino é congruente com os obtidos para *Eimeria* de galinha. A Figura 6.6 mostra que um conjunto de treinamento com 40% (867 imagens) do total de imagens resulta em uma taxa de acerto aceitável. Uma vez que o número total de espécies é de onze, conclui-se que o número mínimo de elementos de treinamento por espécie é de cerca de 80, um valor semelhante ao necessário para *Eimeria* de galinha.

6.3.2 Seleção de características

A Tabela 6.12 apresenta os valores de correlação das distintas características extraídas das imagens de *Eimeria* de coelho. Comparando-se esses dados com os da correlação das características de *Eimeria* de galinha (Tabela 6.2), verifica-se que as correlações entre variáveis de forma geral se mantêm. No entanto, os valores de correlação em *Eimeria* de coelho (Tabela 6.12) são menores do que os observados em *Eimeria* de galinha (Tabela 6.2). Essa concordância, leva a sugerir que as características $F1$, $F8$, $F9$, $F12$ e $F13$ podem ser eliminadas por apresentar uma alta correlação.

Seleção de características usando o classificador por similaridade como função-critério

A Tabela 6.13 apresenta os resultados de seleção de características para as espécies de coelho, usando-se a metodologia SFS e tendo o classificador por similaridade como função-critério.

Tabela 6.12: Média das matrizes de correlação das onze espécies de *Eimeria* de coelho

Nível	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
F1	1,00													
F2	0,02	1,00												
F3	-0,39	0,26	1,00											
F4	-0,81	0,05	0,26	1,00										
F5	-0,57	-0,20	0,33	0,41	1,00									
F6	0,02	0,16	0,05	0,00	-0,13	1,00								
F7	0,02	0,14	0,00	0,06	-0,10	0,09	1,00							
F8	-0,83	-0,07	0,34	0,84	0,81	-0,07	-0,02	1,00						
F9	0,13	0,06	-0,03	-0,14	-0,12	-0,01	0,01	-0,15	1,00					
F10	-0,17	-0,02	0,03	0,18	0,11	0,03	0,01	0,18	-0,81	1,00				
F11	0,27	0,08	0,04	-0,26	-0,23	0,02	0,03	-0,28	0,19	-0,39	1,00			
F12	-0,18	0,01	0,06	0,18	0,08	0,06	0,01	0,15	-0,44	0,76	-0,64	1,00		
F13	0,14	0,02	-0,04	-0,14	-0,08	-0,04	0,00	-0,13	0,83	-0,88	0,52	-0,81	1,00	
F14	-0,17	-0,08	-0,08	0,49	-0,56	0,12	0,16	-0,01	0,00	0,05	0,00	0,09	-0,05	1,00

Figura 6.6: Efeito do tamanho do conjunto de treino na acurácia da classificação em *Eimeria* de coelho. Um total de 2167 imagens foram usadas na avaliação. O tamanho do conjunto de treino é representado pela porcentagem relativa ao total do conjunto de dados. O número absoluto de imagens também é apresentado (em parênteses). Os resultados estão apresentados para o classificador por similaridade (linha cheia) e probabilístico (linha tracejada).

Pode-se verificar que as duas características de maior discriminação correspondem à *F5* e *F8* com 65,95%. Comparando-se esta tabela com a correspondente obtida para *Eimeria* de galinha (Tabela 6.3), observa-se uma similaridade no padrão de discriminação das diferentes características. Embora a maior taxa de acerto seja atingida com 9 características, a diferença do acerto utilizando as melhores combinações entre 7 a 14 características é muito pequena. Um dado interessante é que a característica *F8* (área) é altamente discriminativa para *Eimeria* de coelho, ao passo que para *Eimeria* de galinha esta característica foi pouco discriminativa (Tabela 6.3). Em ambos os conjuntos de parasitas (Tabelas 6.3 e 6.13) verifica-se que as características relativas ao tamanho são as que apresentam o maior poder discriminativo.

Seleção de características usando o classificador por probabilidade como função-critério

A Tabela 6.14 apresenta os resultados obtidos na seleção de características usando SFS, e tendo como função-critério o classificador por probabilidade. Ao contrário dos resultados da Tabela 6.13, a combinação das duas características *F1* e *F2*, resulta na mais discriminativa (61,60%). A maior taxa de acerto (74,90%) foi obtida com a combinação de dez características. Na Tabela 6.14 ganham

Tabela 6.13: Seleção de características usando SFS e o classificador por similaridade para as onze espécies de *Eimeria* de coelho.

Nível	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	Acerto
2					×			×							65,95
3					×		×	×							75,64
4				×	×		×	×							79,74
5				×	×		×	×			×				80,20
6				×	×		×	×		×	×				80,68
7				×	×		×	×		×	×		×		81,78
8				×	×		×	×		×	×	×	×		82,48
9				×	×		×	×	×	×	×	×	×		82,58
10				×	×	×	×	×	×	×	×	×	×		82,09
11	×			×	×	×	×	×	×	×	×	×	×		81,69
12	×	×		×	×	×	×	×	×	×	×	×	×		81,42
13	×	×	×	×	×	×	×	×	×	×	×	×	×		81,81
14	×	×	×	×	×	×	×	×	×	×	×	×	×	×	81,26

Tabela 6.14: Seleção de características usando SFS e o classificador por probabilidade para as onze espécies de *Eimeria* de coelho.

Nível	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	Acerto
2	×	×													61,60
3	×	×					×								70,32
4	×	×					×							×	72,47
5	×	×	×				×							×	73,43
6	×	×	×				×					×	×		73,74
7	×	×	×				×		×			×	×		73,66
8	×	×	×				×		×		×		×	×	74,38
9	×	×	×			×	×		×		×		×	×	74,30
10	×	×	×			×	×	×	×		×		×	×	74,90
11	×	×	×			×	×	×	×		×	×	×	×	74,41
12	×	×	×		×	×	×	×	×		×	×	×	×	74,12
13	×	×	×		×	×	×	×	×	×	×	×	×	×	73,98
14	×	×	×	×	×	×	×	×	×	×	×	×	×	×	73,24

destaque as características de representação da curvatura ($F1$, $F2$ e $F3$). Isso pode ser explicado, pelo menos em parte, pelo fato de haver maior complexidade de formas entre os oocistos de *Eimeria* de coelho do que entre as espécies que infectam a galinha. As espécies de *Eimeria* de galinha basicamente têm três tipos de curvatura: elíptica, ovóide e circular (Figura 1.2). Nas espécies que infectam o coelho, por outro lado, além desses formatos de curvaturas, tem-se oocistos piriformes. Além disso, algumas espécies (*E. intestinalis*, *E. coecicola*, *E. flavescens*, *E. magna* e *E. piriformis* – Veja Figura 6.5) apresentam uma estrutura denominada micrópilo, que constitui uma região achatada num dos pólos do oocisto. Assim, a combinação de um número maior de formatos de curvatura,

associada à existência de micrópilo em algumas espécies, explica porque a curvatura tem uma importância muito maior na discriminação de espécies de *Eimeria* de coelho do que nas espécies de galinha.

6.3.3 Classificação de espécies

As Tabelas 6.15 e 6.16 apresentam as matrizes de confusão para as espécies de *Eimeria* de coelho, utilizando-se os classificadores por similaridade e por probabilidade, respectivamente. Semelhantemente ao que foi observado para *Eimeria* de galinha, o classificador por similaridade apresentou uma taxa de acerto global (80,16%) ligeiramente maior do que o por probabilidade (73,24%). Espécies com morfologia muito discrepante das demais como, por exemplo, *E. exigua* e *E. perforans*, apresentaram as melhores taxas de acerto. Por outro lado, espécies com morfologia claramente semelhante (vide Figura 6.5), como *E. media*, *E. coecicola* e *E. vejovskyi* apresentaram valores nitidamente inferiores.

6.3.4 Desempenho dos classificadores

Da mesma forma como foi verificado para as espécies de *Eimeria* de galinha, além do cálculo das taxas de acerto, também foram geradas as curvas ROC para cada uma das espécies de *Eimeria* de coelho. Considerando as curvas ROC multi-classe, o grafo de cada espécie foi gerado considerando como verdadeiros positivos todas aquelas imagens classificadas corretamente na espécie em análise, e como falsos positivos, as demais imagens classificadas erradamente na espécie em análise.

Analisando a Figura 6.7, verifica-se nas curvas ROC que o classificador por probabilidade (Cp) apresenta um melhor desempenho do que o classificador por similaridade (Cs), o que ratifica os resultados obtidos com as espécies de *Eimeria* de galinha (Figura 6.4). De forma semelhante, o classificador Cs sempre apresentou melhores taxas de acerto do que Cp nas matrizes de classificação.

Tabela 6.15: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. de coelho utilizando-se um classificador por similaridade e validação pelo método “leave-one-out”. Média da diagonal = 80, 16%.

Espécies	Nro. oocistos	Espécies de oocistos										
		<i>E.coe</i>	<i>E.exi</i>	<i>E.fla</i>	<i>E.int</i>	<i>E.irr</i>	<i>E.mag</i>	<i>E.med</i>	<i>E.per</i>	<i>E.pir</i>	<i>E.sti</i>	<i>E.vej</i>
<i>E. coecicola</i>	191	52,88	0,00	0,00	0,00	0,52	0,00	3,14	0,00	1,05	12,04	30,37
<i>E. exigua</i>	282	0,00	99,65	0,00	0,00	0,00	0,00	0,00	0,35	0,00	0,00	0,00
<i>E. flavescens</i>	186	0,54	0,00	91,94	0,00	0,54	0,54	0,00	0,00	0,00	3,23	3,23
<i>E. intestinalis</i>	127	0,79	0,00	0,00	76,38	0,00	0,00	2,36	0,00	18,90	0,00	1,57
<i>E. irresidua</i>	209	0,00	0,00	0,00	0,00	98,09	0,00	0,00	0,00	0,00	1,91	0,00
<i>E. magna</i>	291	0,69	0,00	15,46	0,00	0,69	73,88	0,69	0,00	0,00	8,59	0,00
<i>E. media</i>	199	3,52	0,00	0,00	3,02	0,00	0,50	46,73	3,02	9,05	4,02	30,15
<i>E. perforans</i>	110	0,00	0,00	0,00	0,00	0,00	0,00	2,73	97,27	0,00	0,00	0,00
<i>E. piriformis</i>	133	0,00	0,00	0,00	9,02	0,00	0,75	4,51	0,00	83,46	1,50	0,75
<i>E. stiedai</i>	156	3,21	0,00	1,28	0,00	2,56	0,64	0,00	0,00	0,64	75,64	16,03
<i>E. vej dovskiyi</i>	283	4,59	0,00	0,71	0,35	0,35	0,00	3,89	0,00	0,71	3,53	85,87

Tabela 6.16: Matriz de confusão de diferenciação de espécies de *Eimeria* spp. da coelho utilizando-se o classificador por probabilidade e validação pelo método “leave-one-out”. Média da diagonal = 73,24%.

Espécies	Nro. oocistos	Espécies de oocistos											
		<i>E.coe</i>	<i>E.exi</i>	<i>E.fla</i>	<i>E.int</i>	<i>E.irr</i>	<i>E.mag</i>	<i>E.med</i>	<i>E.per</i>	<i>E.pir</i>	<i>E.sti</i>	<i>E.vej</i>	
<i>E. coecicola</i>	191	53,93	0,00	0,00	1,05	3,14	0,00	11,52	0,00	0,52	15,71	14,14	
<i>E. exigua</i>	282	0,00	100,0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
<i>E. flavescens</i>	186	0,00	0,00	78,26	0,54	1,09	10,87	0,00	0,00	3,80	3,26	2,17	
<i>E. intestinalis</i>	127	1,57	0,00	0,00	66,14	0,00	0,00	16,54	0,00	15,75	0,00	0,00	
<i>E. irresidua</i>	209	0,48	0,00	0,48	0,00	92,34	2,39	0,00	0,00	0,00	4,31	0,00	
<i>E. magna</i>	291	0,00	0,00	5,86	0,00	1,72	85,17	0,00	0,00	0,34	6,90	0,00	
<i>E. media</i>	199	4,57	0,00	0,00	19,29	0,00	0,00	55,84	3,55	10,15	2,03	4,57	
<i>E. perforans</i>	110	0,00	0,00	0,00	0,00	0,00	0,00	10,09	89,91	0,00	0,00	0,00	
<i>E. piriformis</i>	133	0,00	0,00	1,50	23,31	0,00	2,26	8,27	0,00	60,90	3,01	0,75	
<i>E. stiedai</i>	156	8,97	0,00	2,56	0,00	4,49	4,49	0,64	0,00	3,21	63,46	12,18	
<i>E. vejdoskyi</i>	283	10,25	0,00	0,00	1,41	0,35	0,00	10,95	0,00	2,12	15,19	59,72

Figura 6.7: Avaliação comparativa dos classificadores por similaridade (linha azul) e por probabilidade (linha vermelha) através de curvas ROC. A linha preta indica a diagonal de referência. Espécies: (a) *E. coecicola*, (b) *E. exigua*, (c) *E. flavescens*, (d) *E. intestinalis*, (e) *E. magna*, (f) *E. media*, (g) *E. perforans*, (h) *E. piriformis*, (i) *E. stiedai* e (k) *E. vej dovskyi*.

6.4 Sistema de diagnóstico em tempo real

Como prova de conceito (*proof of principle*) de que a abordagem apresentada nesta tese poderia ser aplicada para a discriminação morfológica de espécies de *Eimeria* em um sistema em tempo real para o diagnóstico automático, foi criado o sistema COCCIMORPH. O programa apresenta uma interface *web* (disponível no endereço <http://puma.icb.usp.br/coccimorph>) que permite ao usuário enviar uma imagem via *web* para ser classificada. A interface interativa permite ao usuário escolher o melhor valor de limiar para a etapa de segmentação e observar em tempo real o seu efeito. Uma vez feita a segmentação, o usuário solicita a classificação, a qual é realizada pelo programa e retornada ao usuário também através da interface *web*.

6.4.1 Projeto do sistema de diagnóstico em tempo real

A Figura 6.8 apresenta o projeto do sistema integrado COCCIMORPH, o qual está dividido em três níveis:

Figura 6.8: Fluxo de trabalho do sistema integrado de diagnóstico em tempo real (COCCIMORPH).

- *Banco de Dados*. Neste nível são guardados os vetores de características que compõem o conjunto de dados. As imagens de oocistos isolados também são armazenadas, constituindo o banco de imagens de *Eimeria* (*The Eimeria Image Database*), o qual pode ser visualizado através de uma interface *web* no endereço <http://puma.icb.usp.br/imagedb/>.
- *Aplicação*. Este nível corresponde ao de implementação do sistema, o qual está dividido em três módulos: subsistema de importação, subsistema de análise e módulo de aplicação, e servidor *web*.
- *Cliente*. Este nível está orientado para interagir com o usuário final, permitindo a visualização e envio de imagens para fins de diagnóstico em tempo real.

O subsistema de análise constitui-se no núcleo do sistema, e é responsável pelo pré-processamento da imagem, a extração de características e a classificação de padrões¹. Cada um desses módulos é invocado de forma sequencial e, em conjunto, trabalham de forma integrada com os distintos níveis do sistema, extraindo e armazenando informação no banco de dados, e mostrando os resultados no nível cliente.

O sistema foi desenvolvido em *C++*, o que permite um processamento rápido, possibilitando a implementação do sistema em tempo real via *web*.

6.4.2 Interface *web*

O sistema foi desenvolvido de modo a permitir que os usuários tenham a possibilidade de realizar o diagnóstico de espécies de *Eimeria* de forma remota. A Figura 6.9(a) mostra uma captura da tela principal, com a imagem de um oocisto sendo pré-processado, com o contorno já detectado. Nesta etapa, para se obter a classificação, o usuário somente deve pressionar com o mouse o botão “Classify”, para, em seguida o sistema apresentar uma nova tela com os resultados do diagnóstico (Figura 6.9(b))².

Considerando-se que diferentes usuários têm distintos aparelhos de microscopia ou câmeras digitais, o grau de aumento e a resolução das imagens capturadas podem variar significativamente em relação às imagens usadas neste trabalho. No intuito de normalizar a escala das imagens, a primeira

¹Pode-se verificar que essa sequência de processos corresponde a uma parte da organização desta tese, mostrada no esquema inicial, sendo que cada uma delas forma um capítulo, assim: pré-processamento da imagem (Capítulo 2), extração de características (Capítulo 3) e classificação de padrões (Capítulo 4).

²Para maiores detalhes sobre o uso do sistema, o leitor pode acessar a página do sistema (<http://puma.icb.usp.br/coccimorph>) e consultar a documentação. Essa página também oferece um conjunto de dados exemplo, que permite ao usuário testar o sistema, assim como, seguir os tutoriais.

Figura 6.9: Interface do sistema de diagnóstico em tempo real (COCCIMORPH). (a) Tela principal com imagem de um oocisto enviada ao sistema. O contorno, definido por linha branca circundando o oocisto, pode ser visto. Ao se pressionar o botão *Classifi*, o sistema processa a imagem e gera o resultado da classificação em uma nova tela (b).

tarefa que o usuário deve fazer é determinar o número de *pixels*/ μm das suas imagens capturadas. Isso pode ser facilmente calculado usando-se uma lâmina de microscopia contendo uma escala microscópica calibrada. Alternativamente, também pode ser usada uma câmera de contagem de células do tipo Neubauer (hemocitômetro), de uso comum na grande maioria dos laboratórios biológicos. Uma vez que a escala da imagem tenha sido obtida, a resolução espacial de captura (expressa em *pixels*/ μm) pode ser facilmente determinada usando-se qualquer programa de tratamento de imagens (ex: *Gimp*, *Adobe Photoshop*®, entre outros). Se o usuário capturar todas as imagens subsequentes sob as mesmas condições, esse passo somente deverá ser realizado uma única vez. A interface do sistema COCCIMORPH apresenta uma caixa de formulário em branco (*pixel/micrometer*) na qual o usuário pode colocar o valor da escala das imagens capturadas em seu sistema. O sistema COCCIMORPH normaliza automaticamente a resolução de acordo com aquela utilizada na captura das imagens do banco de imagens de treinamento. Além disso, o sistema também consegue lidar com diferenças de iluminação e contraste, conforme descrito no item 2.4.2.

6.4.3 Banco de imagens

O acesso a uma grande quantidade de dados biológicos permitiu capturar um número grande de imagens de oocistos das diferentes espécies de *Eimeria* de galinha e de coelho. Além de utilizar estas imagens para constituir o banco de treinamento do sistema COCCIMORPH, decidimos também criar um banco público dessas imagens para uso da comunidade científica. Assim, foi criado o sítio *web* denominado *The Eimeira Image Database* (*The Eimeira Image Database*), constituído por amostras de cada uma das espécies. No caso das amostras de *Eimeria* de galinha, também foram incluídas imagens de várias cepas de cada espécie, originadas de diferentes regiões geográficas. A Figura 6.10 apresenta uma tela capturada, mostrando um conjunto de imagens da cepa H de *E. praecox*. No lado esquerdo da tela está disponível um menu para navegar pelas demais espécies. O menu da parte superior da tela permite mudar para outras imagens de micrografias e de *Eimeria* de coelho. Esse banco de imagens poderá ser muito útil para o treinamento de pessoal especializado em diagnóstico tendo, assim, um caráter didático. Além disso, por ser constituído por imagens de amostras puras de cada espécie, o banco representa um “*golden standard*” para o desenvolvimento de novos algoritmos de extração de características, bem como de novos classificadores, cujos resultados poderiam ser confrontados com os descritos neste trabalho.

Figura 6.10: Tela capturada do sítio web do “*Eimeria Image Database*”, apresentando imagens de oocistos isolados da cepa H de *E. praecox*.

6.4.4 Morfometria dos oocistos

A partir dos dados morfométricos, foi feita uma revisão das medidas de tamanho dos oocistos de *Eimeria* de galinha (Tabela 6.17). As medidas obtidas estão em concordância com Conway e McKenzie (1991) assim como com os relatados no trabalho de Kucera e Reznicky (1991). Verifica-se que através do valor de razão (D/d), também conhecida como *shape index*, a espécie mais próxima da unidade é a mais circular (*E. mitis*), e a mais alongada (elíptica) aquela de maior valor (*E. acervulina*).

Tabela 6.17: Dados morfométricos de *Eimeria* spp. de galinha doméstica. Os valores mínimos (Mín) e máximos (Máx) dos diâmetros maior (D) e menor (d), assim como os respectivos desvios padrão (D.P.) estão apresentados. O *shape index* (Razão D/d) também está apresentado.

Espécie	# de oocistos	Diâmetro maior (D) μm				Diâmetro menor (d) μm				Razão D/d
		Mín	Máx	Média	D.P.	Mín	Máx	Média	D.P.	
<i>E. acervulina</i>	374	15,4	24,1	19,0	1,29	11,5	17,1	13,8	0,97	1,38
<i>E. brunetti</i>	418	21,8	30,4	26,4	1,50	17,5	24,4	21,1	1,14	1,25
<i>E. maxima</i>	103	28,2	36,2	32,2	1,59	22,0	29,9	25,3	1,47	1,27
<i>E. mitis</i>	335	11,9	20,6	16,8	1,41	11,3	19,0	15,4	1,21	1,09
<i>E. necatrix</i>	259	17,0	27,6	21,2	1,64	13,0	19,5	16,7	1,06	1,27
<i>E. praecox</i>	377	15,9	25,2	20,0	1,49	14,9	21,9	17,7	1,16	1,13
<i>E. tenella</i>	311	17,3	24,5	21,5	1,23	15,2	20,8	18,0	0,91	1,20

Os dados morfométricos das onze espécies de *Eimeria* de coelho doméstico (Tabela 6.18) também foram determinados. A espécie *E. exigua* apresentou-se como a menor e mais circular, enquanto a espécie *E. irresidua* foi a maior de todas as espécies e *E. coecicola*, *E. vej dovskyi* e *E. stiedai* como as mais alongadas.

Tabela 6.18: Dados morfométricos de *Eimeria* spp. de coelho doméstico. Os valores mínimos (Mín) e máximos (Máx) dos diâmetros maior (D) e menor (d), assim como os respectivos desvios padrão (D.P.) estão apresentados. O *shape index* (Razão D/d) também está apresentado.

Espécie	# de oocistos	Diâmetro maior (D) μm				Diâmetro menor (d) μm				Razão D/d
		Mín	Máx	Média	D.P.	Mín	Máx	Média	D.P.	
<i>E. coecicola</i>	191	31,1	41,6	35,4	1,81	17,3	23,8	20,3	1,20	1,75
<i>E. exigua</i>	282	14,0	19,9	16,8	1,16	13,1	18,0	14,9	0,97	1,12
<i>E. flavescens</i>	186	27,2	39,0	33,3	1,94	19,4	26,6	22,6	1,63	1,47
<i>E. intestinalis</i>	127	27,1	34,6	30,7	1,33	18,6	22,4	20,3	0,78	1,51
<i>E. irresidua</i>	209	35,6	44,8	39,4	1,71	20,8	27,5	24,0	1,03	1,64
<i>E. magna</i>	291	31,7	41,4	36,2	1,83	20,7	27,9	24,1	1,21	1,50
<i>E. media</i>	199	25,1	37,6	31,2	1,91	15,6	22,2	19,1	1,34	1,63
<i>E. perforans</i>	110	20,2	27,1	23,7	1,97	13,0	18,1	15,0	1,03	1,58
<i>E. piriformis</i>	133	27,2	35,1	31,6	1,36	17,8	25,2	21,1	1,11	1,50
<i>E. stiedai</i>	156	32,1	39,6	36,0	1,60	19,0	26,0	21,7	1,34	1,66
<i>E. vej dovskyi</i>	283	29,8	38,9	34,5	1,68	17,9	23,1	20,4	0,90	1,69

6.5 Análise de distância entre espécies baseada em dados morfológicos

Os dados morfométricos utilizados no diagnóstico de *Eimeria* também foram usados para se construir árvores de distância entre as distintas espécies de *Eimeria*. Como os dados de morfologia eram contínuos, foi necessário ser fazer uma discretização a fim de se gerar uma matriz de caracteres. Conforme será apresentado a seguir, essas árvores foram comparadas com aquelas geradas por métodos de filogenia molecular os quais usam marcadores filogenéticos clássicos como o gene da subunidade menor (18S) do ribossomo ou genomas mitocondriais

6.5.1 Matriz de caracteres

Seguindo o esquema (Figura 5.1) descrito na Seção 5.4 cada caractere foi discretizado, produzindo um conjunto de intervalos. As médias de medidas de cada espécie foram designadas no seu respectivo intervalo e os números desses intervalos foram usados para compor a matriz de caracteres. Diferentes métodos de discretização foram empregados (item 5.3), de forma a gerar intervalos de mesmo tamanho, mesma frequência ou por entropia. A Tabela 6.19 mostra a matriz de caracteres das espécies de *Eimeria* de galinha obtida por discretização realizada com o método de frequências iguais.

Tabela 6.19: Matriz de caracteres de dados morfométricos

Espécies	A	B	C	D	E	F	G	H	I	J	K	L	M
<i>E.acervulina</i>	5	5	1	2	1	4	4	1	3	2	4	2	4
<i>E.maxima</i>	1	4	5	5	5	3	5	5	3	4	1	5	3
<i>E.brunetti</i>	1	3	5	5	5	3	4	5	3	4	1	4	3
<i>E.mitis</i>	5	1	2	1	2	3	2	1	4	2	5	2	4
<i>E.praecox</i>	3	2	4	3	3	3	2	3	3	3	4	3	3
<i>E.tenella</i>	2	4	3	4	4	4	4	4	3	4	3	3	3
<i>E.necatrix</i>	3	4	3	3	3	4	3	3	3	3	3	3	3

6.5.2 Árvores de distância e inferencia filogenética

A matriz de caracteres (Tabela 6.19) constitui-se na matriz de entrada para os programas de inferência filogenética. Neste trabalho foi usado o programa de uso público de acesso *on-line* da Universidade de Alberta, para o cálculo de clustering, disponibilizado na página *web*:

<http://www2.biology.ualberta.ca/jbrzusto/cluster.php>

6.5. ANÁLISE DE DISTÂNCIA ENTRE ESPÉCIES BASEADA EM DADOS MORFOLÓGICOS 99

Os parâmetros usados no programa, tendo como entrada nossa matriz de caracteres, foram os seguintes: *Input data is: Character Matrix: samples are rows; Location of Sample Names: Data Windows; Distance/Similarity Measure: Canberra Distance; Clustering Method: Saitou and Nei Neighbour Joining; Stability Analysis: Jackknife; Support Type: Traditional: group=all descend*

A árvore resultante está mostrada na Figura 6.11(a) onde verifica-se a formação de dois clados principais. O primeiro clado é formado pelas espécies *E. acervulina*, *E. mitis*, e o segundo pelas espécies *E. brunetti* e *E. maxima*. *E. praecox* apresentou-se mais próxima de *E. mitis* e *E. acervulina*, enquanto *E. tenella* ficou topologicamente mais próxima de *E. brunetti* e *E. maxima*.

Figura 6.11: Cladogramas gerados (a) com dados morfométricos e (b) por inferência filogenética utilizando-se genomas mitocondriais completos com máxima verossimilhança e o modelo de substituição de nucleotídeos GTR+I+G.

Ao se comparar a árvore obtida com caracteres morfológicos (Figura 6.11(a)) com a obtida por Romano (2004) através de dados moleculares (Figura 6.11(b)), verifica-se uma alta similaridade na topologia das árvores. De fato, em ambos os métodos as espécies *E. acervulina* e *E. mitis*, e *E. brunetti* e *E. maxima*, estão nos mesmos clados, respectivamente. Esses dados também são compatíveis com a topologia obtida por Barta et al. (1997) a partir de seqüências nucleotídicas do gene 18S do citron ribossômico. As árvores de genoma mitocondrial e de 18S apresentam as espécies *E. tenella* e *E. necatrix* formando um único clado, o que não se mostra em nosso resultado da Figura

6.11(a), embora essas espécies estejam muito próximas na árvore obtida por dados morfométricos. De fato, algumas das árvores obtidas por diferentes métodos de discretização também apresentaram esse clado (dados não mostrados). De forma geral, pode-se concluir que árvores de distância obtidas por dados morfométricos revelaram topologias bastante similares às observadas por métodos clássicos de inferência filogenética usando marcadores moleculares.

Capítulo 7

Discussão e conclusões

7.1 Discussão

7.1.1 Resumo do trabalho

Nesta tese descreveu-se o desenvolvimento de uma abordagem integrada de reconhecimento de padrões aplicada à caracterização morfológica e discriminação automática de diferentes espécies de parasitas protozoários do gênero *Eimeria*. Foi proposto o uso de um conjunto de características que envolvem três categorias: (1) medidas geométricas, (2) caracterização da curvatura e (3) quantificação da estrutura interna. Essas características foram extraídas automaticamente e utilizadas para compor o vetor de características de dimensão 14, o qual representa uma versão reduzida da imagem. Os padrões obtidos foram classificados usando-se duas abordagens de discriminação Bayesiana: a primeira usa como função de verossimilhança a Gaussiana e fornece resultados de similaridade, enquanto a segunda usa a função de Dirichlet e fornece resultados probabilísticos.

O trabalho foi desenvolvido e padronizado usando-se imagens microscópicas obtidas a partir de amostras puras de cada uma das espécies do parasita. Um grande número de imagens, constituindo no total 3891 micrografias de oocistos, foi usado para reduzir o efeito da heterogeneidade da morfologia. Adicionalmente, sempre que possível, foram usadas várias cepas de cada espécie, coletadas de diferentes regiões geográficas, no intuito de diluir possíveis variações intra-específicas e maximizar a discriminação inter-específica. Outras fontes de variabilidade dos dados também foram identificadas e avaliadas, incluindo diferenças no contraste e na iluminação do microscópio, assim como o volume da suspensão de parasitas entre a lâmina e a lamínula. Finalmente, foi usado um número relativamente alto de características, as quais foram submetidas a um processo de seleção

para avaliar quantas e quais delas poderiam constituir o conjunto mais discriminante.

A abordagem descrita neste trabalho mostrou-se simples e robusta, permitindo uma identificação confiável das diferentes espécies de parasitas. As características não estão somente limitadas às mais simples e tradicionais medidas geométricas, mas também foi computada a curvatura de cada objeto para representar a forma, e o usadas técnicas de extração de textura para a caracterização da estrutura interna. Foram obtidas taxas de discriminação correta de espécies de aproximadamente 85% (com o classificador por similaridade) e 80% (como o classificador por probabilidade). Esses resultados podem ser considerados muito bons, especialmente se considerarmos que o diagnóstico é baseado em morfologia, a qual sempre apresenta certo grau de sobreposição. Contudo, ao contrário do diagnóstico obtido por inspeção visual humana, não há subjetivismo nessa discriminação. Além disso, dada a complexidade dos algoritmos para extração de características, a implementação do sistema é computacionalmente eficiente, permitindo uma rápida interação em tempo real do usuário final através de uma interface *web*.

Considerando-se que a abordagem usa algoritmos genéricos, o sistema pode ser facilmente ampliado para a discriminação de outros organismos, desde que os mesmos apresentem uma morfologia semelhante. Para essa tarefa, o usuário só precisa fornecer de um novo banco de imagens e usá-los no treinamento do sistema para a discriminação das diferentes classes. De fato, os resultados preliminares, mostrados com as onze espécies de *Eimeria* que infectam o coelho doméstico, mostraram um desempenho discriminante similar. Por outro lado, no caso de organismos cujo domínio de imagens seja muito diferente, novas outras características discriminantes teriam de ser encontradas. Uma vez feito isso, entretanto, os mesmos classificadores e implementação de interface *web* poderiam ser empregados.

7.1.2 Comparação do sistema COCCIMORPH com outros trabalhos

Estudos prévios usando processamento digital de imagens, aplicados a *Eimeria* (Kucera e Reznicky, 1991; Dauschies et al., 1999; Plitt et al., 1999) foram descritos na literatura. Esses sistemas, no entanto, estão restritos a um procedimento semi-automático de cálculo dos diâmetros dos oocistos e, ainda, requerem uma forte interferência humana durante o processamento. Além do mais, muitos dos estudos empregaram um baixo número de características morfológicas. Assim, alguns trabalhos usaram como características unicamente os diâmetros dos oocistos (Kucera e Reznicky, 1991; Plitt et al., 1999), enquanto que outros usaram a transformada de Fourier do contorno (Sommer, 1998a) ou calcularam algumas estatísticas sobre a transformação feita (Yang et al., 2001).

Outra limitação geral esta relacionada ao método de classificação, onde as distribuições de dados

multi-dimensionais não foram consideradas. Sommer (1998a) usou a distancia Euclideana como métrica para a aglomeração (*clustering*). Essa métrica assume que os dados estão distribuídos de forma homogênea, o que não é necessariamente o caso, especialmente quando são usados dados multi-dimensionais. Yang et al. (2001), trabalhando com ovos de helmintos humanos, usou quatro características morfométricas e duas camadas de redes neurais artificiais (ANNs). Essas ANNs foram usadas para a identificação dos ovos e sua respectiva diferenciação dos artefatos, e para a discriminação de espécies, respectivamente. A estimativa da taxa média de classificação correta foi, portanto, baseada em um pequeno número de imagens, e o possível efeito da variabilidade intra-específica não foi avaliada pelos autores.

Em relação a resultados obtidos por outros autores, Kucera e Reznicky (1991) não conseguiram diferenciar totalmente as espécies de *Eimeria*, uma vez que esses autores somente utilizaram os diâmetros maior e menor, e realizaram a diferenciação com base na distribuição de histogramas dos diâmetros dos oocistos. Por outro lado, os resultados obtidos por Dauschies et al. (1999), que trabalhou com *Eimeria* de suínos, mostraram uma alta taxa de classificação correta (99%), mas o procedimento empregado não foi totalmente automático e as classes usadas não foram balanceadas. De fato, os parasitas foram manualmente divididos, previamente ao processamento, em três grupos, de acordo com a rugosidade da camada externa da parede dos oocistos. Uma tentativa de automatizar o procedimento proposto por Dauschies et al. (1999), foi apresentado por Plitt et al. (1999), obtendo-se uma alta taxa de classificação (97%), mas o procedimento de extração de características morfológicas continuou a ser semi-automático. Yang et al. (2001), usando ovos de helmintos de humanos, conseguiram uma taxa de acerto de 83%, mas esses resultados foram obtidos com uma quantidade muito pequena de elementos teste.

Os resultados de classificação correta obtidos em nosso trabalho de 85% e 80%, com o uso dos classificadores por similaridade e por probabilidade, respectivamente, mostram-se muito bons em relação àqueles descritos na literatura. Além disso, e ao contrário do que é apresentado na literatura, o nosso sistema fornece um maior corpo de evidências para que o usuário tome a decisão em relação ao diagnóstico. Assim, nosso sistema utiliza um total de 14 características distintas, além de dois classificadores. Os resultados não se limitam a um simples diagnóstico, mas, sim, a uma lista de escores de similaridade, com imagens protótipo de cada espécie, ou de probabilidade. Assim, nos casos limítrofes (*borderline*), o usuário não é simplesmente confrontado com um resultado simples e único, fruto de uma decisão automática do programa. Ao contrário, o usuário, ao receber a lista de valores de similaridade e probabilidade, pode tomar uma decisão diagnóstica mais consciente, sabendo, inclusive, que outra(s) espécie(s) poderia(m) ser a correta para o caso.

7.1.3 Aquisição de imagens

Alguns problemas apresentados na aquisição de imagens foram sendo gradativamente superados ao longo do trabalho, levando a um protocolo altamente padronizado para este procedimento. Por exemplo, a utilização de um volume muito pequeno (cerca de $3 \mu l$) de suspensão de oocistos entre a lâmina e lamínula levava a uma alteração da morfologia dos oocistos. Assim, devido ao pequeno volume da suspensão, os oocistos sofriam maior pressão do peso da lamínula, fazendo com que houvesse um achatamento dos mesmos e o aumento artefactual dos seus diâmetros. Por outro lado, o uso de volumes muito maiores resultava em oocistos muito “soltos” no campo microscópico, além de reduzir o contraste das micrografias. Um bom compromisso foi conseguido com o volume de $6 \mu l$, o qual já reduzia o peso da lamínula devido à ação da alta tensão superficial da água e, ao mesmo, tempo, mantinha um alto contraste para a captura de imagens.

Outro aspecto importante diz respeito ao ajuste do foco no microscópio. Como o oocisto é uma estrutura bastante grande, não é possível se ter no microscópio uma profundidade de campo grande o suficiente para manter em foco o oocisto ao longo de toda a sua espessura. Assim, deve-se escolher um plano focal que coincida com o eixo do diâmetro maior do oocisto. Esse ponto pode ser facilmente identificado, pois é o plano focal no qual a parede do oocisto aparece com aspecto mais fino. Saindo-se desse plano focal, há um aumento perceptível da espessura aparente da parede.

Outro parâmetro importante na captura é a iluminação do microscópio, a qual depende da intensidade da lâmpada, do nível de abertura do diafragma, da posição do condensador e do uso de filtros. Em nosso trabalho procuramos padronizar ao máximo as condições de captura para reduzir variabilidades. Entretanto, como nosso sistema foi concebido para lidar com imagens de usuários remotos, os quais podem estar utilizando equipamentos de microscopia e captura de imagem muito diferentes, foi necessário se criar um mecanismo de normalização. Assim, a partir do conjunto do treinamento, foi determinada uma imagem virtual que denominamos de protótipo, ou seja, um elemento modelo que apresenta valores numéricos médios para cada uma das diferentes características morfológicas. O histograma de distribuição de níveis de cinza foi calculado para este protótipo e usado para se normalizar as imagens consulta. Assim, quando uma imagem consulta é enviada ao sistema, seu histograma é determinado e equalizado com base no histograma da imagem protótipo, conforme metodologia descrita no item 2.4.2. Com isso, consegue-se reduzir, dentro de certos limites, o efeito de diferenças de iluminação na captura.

Outro obstáculo para que o sistema pudesse ser utilizado por usuários remotos era relacionado com a resolução de captura. Assim, como a morfometria é baseada na quantificação de *pixels*, alterações na resolução de captura resultam em medidas que podem ser totalmente diferentes,

ainda que o objeto capturado seja o mesmo. Para que usuários com câmeras digitais de diferentes resoluções pudessem utilizar o nosso sistema, foi criado um campo na interface *web* no qual pode-se definir a resolução linear da imagem (escala), medida em *pixels* por micrômetro. Assim, um usuário determina a resolução linear de captura de seu equipamento de microscopia e fotografia digital, e informa o COCCIMORPH. O nosso sistema então normaliza a resolução da imagem consulta segundo a escala definida pelo usuário, adequando-a à resolução das imagens do conjunto de treinamento. É evidente que essa abordagem é limitada pelo mínimo de informação necessário para que o sistema discrimine corretamente a imagem. Em nossos testes, a menor resolução que ainda permitiu discriminar adequadamente os oocistos foi a de 1280 x 960 (ver Tabela 2.2). Abaixo dessa resolução o sistema foi capaz de discriminar apenas uma fração das imagens. Contudo, com a recente popularização das câmeras digitais e a queda vertiginosa dos preços, dificilmente um usuário irá utilizar atualmente uma câmera com resolução inferior a 5 *megapixels*.

7.1.4 Das características extraídas

Inicialmente foram testadas características que deveriam identificar a morfologia dos oocistos como aquelas relacionadas à curvatura e às medidas geométricas. Embora os resultados iniciais tenham mostrado uma boa diferenciação, a adição de características para representar a estrutura interna ajudou a incrementar a taxa de discriminação. Contudo, a determinação de quais características e sua combinação é a mais discriminativa, está sujeita aos diversos fatores que influenciam na classificação.

No caso das espécies de *Eimeria* de galinha, as características *F4*, *F5*, *F7* e *F11* revelaram-se as mais discriminativas (Tabelas 6.3 e 6.4) no processo de seleção de características. Isso sugere que o tamanho (*F4* e *F5*), a simetria no eixo menor (*F7*) e o contraste da estrutura interna (*F11*) são os aspectos morfológicos importantes na diferenciação das espécies de *Eimeria* de galinha. Ao se analisar a Tabela 6.3, que usa o classificador por similaridade como função-critério, observa-se que outras características da estrutura interna ganham importância (*F12* e *F13*). Por outro lado, na Tabela 6.4 chama a atenção a importância que ganha a característica *F14*. Nas duas Tabelas, 6.3 e 6.4, as características da curvatura (*F1*, *F2* e *F3*) estão entre as que menos influenciam na discriminação. Isso pode indicar que, nas espécies de *Eimeria* de galinha, a curvatura não é tão importante na sua discriminação, mas, sim, aspectos relacionados ao tamanho, simetria e contraste da estrutura interna.

Nas espécies de *Eimeria* de coelho, verifica-se que quando é usado o classificador por similaridade (Tabela 6.13), as características relacionadas ao tamanho também têm grande importância (*F4*,

F5, *F7* e *F8*), mostrando um papel similar ao observado com as espécies de galinha (Tabelas 6.3 e 6.4). Por outro lado, ao se aplicar o classificador por probabilidade (Tabela 6.14), as características relacionadas à curvatura apresentam-se como as mais discriminativas (*F1*, *F2* e *F3*), juntamente com *F7* e *F14*. Isso pode ser explicado, pelo menos em parte, pelo fato de as espécies de *Eimeria* de coelho apresentarem maior variação de formato de curvatura do que as de galinha. Assim, enquanto as espécies de *Eimeria* de galinha têm formatos circular, ovóide e elíptico, as espécies de coelho apresentam ainda o formato piriforme (*E. piriformis* e *E. intestinalis* - Figura 6.5(c) e (g)). Além disso, algumas espécies (*E. intestinalis*, *E. coecicola*, *E. flavescens*, *E. magna* e *E. piriformis*) também apresentam uma região achatada em um dos pólos do oocisto, denominada micrópilo. Por outro lado, na Tabela 6.12 verifica-se que as características *F2*, *F6* e *F14* não têm correlação com outras. Estas por sua vez, no senso comum, deveriam mostrar-se discriminativas, o que na prática não aconteceu. Analisando as Tabelas 6.3 e 6.13, observa-se que as características com menor correlação são as que apresentaram menor efeito discriminativo. Isso nos leva a concluir que, embora algumas características não apresentem correlação, isso não necessariamente significa que as mesmas sejam discriminativas. Uma outra observação interessante está relacionada à característica *F6* (simetria em função do diâmetro maior), a qual apresenta pouca variabilidade entre as distintas espécies, sendo uma das que têm menor poder discriminativo. Isso significa que os oocistos são muito simétricos quando observados na posição “deitada”.

Um aspecto importante do conjunto de características utilizadas, é que elas se mostram invariantes à translação e rotação do objeto de interesse. Isso significa que a posição e orientação do objeto podem ser diferentes, sem prejuízo da capacidade de discriminação do sistema. Por outro lado, o tamanho, conforme discutido acima, é uma característica altamente discriminativa e, como é baseada em valores absolutos, é altamente variante em função da resolução de captura. Contudo, conforme discutido no item 7.1.3, nosso sistema normaliza a resolução das imagens consulta em função dos protótipos do conjunto de treinamento e, desta forma, consegue-se de maneira artificial tornar o tamanho uma característica praticamente invariante.

7.1.5 Da classificação

Para se escolher o classificador a ser utilizado pelo sistema, foram feitos testes preliminares considerando-se algumas metodologias alternativas de classificação, tais como o SVM (Cristianini e Shawe-Taylor, 2000; Crammer e Singer, 2001). Especificamente, foram comparadas as acurácias dos classificadores Bayesiano e SVM considerando-se situações envolvendo as sete espécies de *Eimeria* de galinha e um conjunto de 14 características. Uma vez que os resultados obtidos não indicaram um desempenho superior da metodologia SVM, decidiu-se usar a metodologia Bayesiana.

De fato, resultados levemente melhores foram conseguidos com o classificador Bayesiano por similaridade. Além disso, uma razão adicional que motivou essa escolha é fato de que o classificador Bayesiano é muito mais simples de ser implementado para um sistema interativo que trabalha em tempo real.

Embora a abordagem Bayesiana forneça resultados probabilísticos, algumas técnicas baseadas na fórmula de Bayes também geram resultados não probabilísticos. Assim, técnicas que usam como função de densidade a normal, por exemplo, definem regiões que por sua vez são representadas através de funções discriminantes (Duda et al., 2001), não alterando a classificação final. Nesse caso, para cada elemento é gerado um escore de classificação. Usando-se a distância de Mahalanobis, pode-se interpretar os resultados como sendo de similaridade da imagem consulta em relação ao protótipo do conjunto de treinamento, daí a denominação atribuída ao classificador (por similaridade).

Por outro lado, o classificador que usa como função de densidade Dirichlet (Carlin e Louis, 1996; Pereira e Stern, 1999, 2001), fornece resultados probabilísticos de classificação e está baseado no cálculo não paramétrico da densidade, daí a sua denominação como classificador por probabilidade.

Para se obter uma alta taxa de acerto, foi preciso inicialmente se determinar o tamanho mínimo do conjunto de treinamento. Essa taxa poderia variar em função da estrutura interna dos classificadores, mas, principalmente, pela escolha das características usadas para a discriminação. Conforme apresentado no item 6.2.3, ambos os classificadores (por similaridade e probabilístico) estabilizam as suas taxas de acerto a partir de cerca de 96 elementos por espécie, o que está em concordância com o relatado por (Jain et al., 2000), que recomenda usar como exemplos de treinamento um número pelo menos dez vezes maior do que o de características. Em nosso trabalho, com *Eimeria* de galinha, obtivemos uma estabilização da taxa de acerto com 9 a 10 características. Resultados similares foram obtidos com as espécies de *Eimeria* de coelho.

Na Figura 6.1 verifica-se que o classificador por similaridade sempre apresenta uma melhor média de taxa de acerto nas distintas proporções de conjunto de treinamento. O mesmo é observado quando são aplicadas distintas estratégias de amostragem do classificador (“Todos contra Todos”, “Partição aleatória” e *Leave One Out*) sendo que, o classificador por similaridade atinge 85% contra 80% do classificador por probabilidade, aproximadamente (Figura 6.2). Analisando-se as matrizes de confusão (Tabelas 6.5 - 6.10), verifica-se que o classificador por similaridade apresenta melhores resultados em todas espécies, exceto em *E. acervulina*. Isso pode sugerir que o classificador por probabilidade é bom para discriminar a espécie *E. acervulina*, e o classificador por similaridade para as demais espécies. Note-se que a diferença na taxa média de acerto para as espécies *E. praecox*, *E. tenella* e *E. necatrix* é significativamente superior com o uso do classificador por similaridade.

O uso da taxa média de acerto (acurácia) não necessariamente garante que o classificador com maior acerto seja o melhor, pois a acurácia assume que a distribuição das classes é constante e que as mesmas estão relativamente balanceadas (Provost e Fawcett, 1997), uma situação que nem sempre se apresenta em problemas do mundo real. Por outro lado, o uso de curvas ROC ajuda a ter uma visão mais geral do desempenho do classificador. Como pode ser visualizado na Figura 6.4, o classificador por probabilidade apresenta curvas melhores do que as do classificador por similaridade, ainda que as taxas médias de acerto foram melhores com o classificador por similaridade. Uma possível interpretação (Fawcett, 2006) desse resultado é de que o classificador por probabilidade é mais “conservador”, isto é, realiza classificações positivas somente com uma alta evidência, produzindo assim poucos falsos positivos. Por outro lado, como consequência desse comportamento, esse classificador também produz um menor número de verdadeiros positivos. O classificador por similaridade, por sua vez, pode ser considerado mais “liberal”, por realizar classificações positivas com pouca evidência, resultando em um maior número de classificações corretas, mas com uma taxa de falsos positivos também maior.

O classificador por probabilidade gera como resultado um conjunto de probabilidades de um dado elemento pertencer a cada uma das classes utilizadas no processo de classificação. Quando a probabilidade desse elemento pertencer a uma classe é muito alta, isto corresponde a dizer que há evidências suficientes para se tomar a decisão de classificá-lo nessa classe. Por outro lado, quando a probabilidade desse elemento pertencer à classe não é muito alta, ainda que seja esta a classe de maior probabilidade, isto significa que as evidências estão distribuídas em outras classes com um menor peso. Por exemplo, seja o vetor de classificação abaixo, cujas probabilidades estão expressas por valores entre 0 e 1:

(0,6; 0,3; 0,1; 0,0; 0,0; 0,0; 0,0)

O fato do elemento apresentar 60% de probabilidade de pertencer à primeira classe significa que existe de fato um conjunto grande de evidências que suportam essa classificação, ainda que existam probabilidades significativas de o elemento pertencer à segunda ou terceira classes.

Por outro lado, analisemos o vetor de probabilidade abaixo, relativo a um segundo elemento:

(0,30; 0,20; 0,10; 0,05; 0,05; 0,15; 0,15)

Embora a primeira classe apresente a maior probabilidade, há uma distribuição de probabilidades relativamente altas nas demais classes, cuja soma é de 70%. Portanto, isso significa que as evidências de suporte à classificação do elemento na primeira classe são bem menores.

Analisado o mesmo elemento sob um classificador por similaridade, os resultados apresentados refletem escores e, portanto, sua soma não é a unidade. Assim, um vetor de similaridade do primeiro

elemento do exemplo acima seria:

(0,85; 0,72; 0,45; 0,1; 0,0; 0,0; 0,0)

Novamente existe uma forte evidência de que o elemento seja pertencente à classe 1. Contudo, como esse elemento também apresenta um alto valor de similaridade com a segunda classe, não se pode ter tanta confiança no resultado quanto no caso do classificador por probabilidade.

Concluindo, o problema de avaliação dos classificadores não é trivial e, portanto, não consiste em simplesmente se determinar qual dos classificadores foi o melhor, mas, sim, em se analisar os resultados obtidos sob distintas perspectivas de classificação. Nesse sentido, os resultados dos dois classificadores oferecem dois pontos de vista diferentes fornecendo maiores evidências para ajudar o usuário a tomar a decisão final. Uma tabela com resultados comparativos dos dois classificadores, usando-se o conjunto de 3891 imagens de *Eimeria*, está disponível publicamente no endereço <http://puma.icb.usp.br/coccimorph/classification/classification.html>.

7.1.6 Um novo conceito em diagnóstico de parasitas

A abordagem integrada apresentada neste trabalho, e totalmente implementada na forma de uma interface *web* que permite uma interação com o usuário, é um novo paradigma na área de diagnóstico de parasitas.

Comparando com outras abordagens de diagnóstico, o sistema apresentado não requer pessoal treinado na identificação de parasitas ou em técnicas de biologia molecular. Além disso, como o sistema pode funcionar de forma remota através da submissão de imagens, não há necessidade de se realizar o transporte físico de amostras biológicas entre a granja e o laboratório de referência. Esse é um aspecto particularmente importante, uma vez que o tráfego de amostras vivas representa um potencial risco sanitário devido à possível disseminação de doenças. Assim, granjas localizadas em áreas com surtos de doenças controladas poderiam enviar amostras digitais para laboratórios de referência sem o risco de espalhar essas doenças em regiões livres das mesmas.

Outra grande vantagem dessa abordagem é que o diagnóstico é obtido em tempo real, de forma imediata, o que possibilita uma tomada de decisão rápida por parte do usuário. Essa característica de resposta imediata é muito importante na área médica, pois pode permitir ao clínico identificar rapidamente a espécie do parasita que está infectando um paciente e, assim, tomar a decisão de que agente terapêutico empregar.

O sistema COCCIMORPH, desenvolvido nesse trabalho para o diagnóstico de parasitas do gênero *Eimeria*, é um modelo de implementação de uma ferramenta de diagnóstico em tempo real

através de uma interface *web*. Usando essa interface, o usuário pode enviar imagens de *Eimeria* para o nosso servidor, segmentá-las *online* e obter uma classificação em tempo real. A única exigência do sistema é de que o usuário deve calibrar o seu sistema de microscopia e captura digital, isto é, determinar a resolução das suas imagens. Uma vez realizado esse processo, e desde que o usuário não utilize outros parâmetros ou equipamentos, todos os diagnósticos poderão ser feitos sem nenhum trabalho adicional.

Do ponto de vista de acurácia, o sistema apresentou taxas de acerto global da ordem de 80-85%, com algumas espécies com taxas relativamente inferiores, de cerca de 70%, e algumas outras próximas de 100%. Esses valores podem ser considerados bastante satisfatórios considerando-se que se trata de uma sistema de diagnóstico baseado em morfologia. Afinal, o objetivo de um sistema como esse não é o de competir com ensaios moleculares como a PCR, mas sim, com diagnósticos morfológicos feitos por inspeção visual. Além disso, deve-se considerar que o usuário não mandaria de forma geral uma única imagem do parasita, mas, sim, uma ou duas dezenas. Assim, possíveis erros de classificação devidos a variações morfológicas mais extremas seriam diluídos, e a taxa de acerto global provavelmente aumentaria substancialmente. Finalmente, é importante mencionar que, ao contrário da inspeção visual, o diagnóstico é realizado sem subjetivismo.

A incorporação de outros parasitas ao sistema poderia também incrementar o escopo de aplicabilidade dessa ferramenta de diagnóstico eletrônico. Protozoários da classe Coccidia e ovos de helmintos, que apresentam uma morfologia similar aos oocistos de *Eimeria*, são os óbvios candidatos a serem incluídos num futuro próximo. De fato, embora o sistema tenha sido todo construído usando-se imagens de *Eimeria* de galinha, a incorporação das onze espécies de *Eimeria* que infectam o coelho foi relativamente fácil e rápida, com resultados de discriminação bastante similares. No caso de parasitas com morfologia muito diferente, seria necessário se padronizar um novo conjunto de características morfológicas que pudesse discriminar adequadamente este novo domínio de imagens. Ainda assim, os classificadores poderiam ser os mesmos relatados aqui, assim como todo o sistema de interface de usuário via *web*.

Com a atual queda nos preços das câmeras digitais de alta resolução (acima de 4,0 *Mpixels*), a metodologia apresentada é relativamente de baixo custo. De fato, qualquer microscópio de qualidade intermediária, acoplado a um sistema de aquisição de imagens digitais (uma câmera e um tubo adaptador) poderia representar o sistema mínimo requerido para tal metodologia. Concluindo, acreditamos que a abordagem proposta neste trabalho demonstra a viabilidade de se utilizar sistemas assistidos por computador como uma interessante alternativa para o diagnóstico de parasitas, de forma rápida, barata e sem riscos sanitários.

7.1.7 Um repositório de imagens e de dados morfométricos de parasitas

O conjunto de milhares de imagens de oocistos de *Eimeria* de galinha e de coelho foi organizado em um sítio *web* e disponibilizado publicamente, o qual foi denominado como *The Eimeria Image Database* (<http://puma.icb.usp.br/imagedb/>). O banco de imagens inclui atualmente cerca de 4500 imagens de *Eimeria* de galinha e aproximadamente 3500 imagens das onze espécies de *Eimeria* que infectam o coelho doméstico. É importante ressaltar que todas as imagens provêm de amostras puras de parasitas, cujo grau de pureza foi monitorado pelo padrão da infecção, características morfológicas e também por PCR. Uma vez que este banco de imagens poderá ser incrementado com imagens de outros parasitas no futuro, ele poderá vir a representar uma valiosa fonte de consulta para parasitologistas clássicos, assim como também para fins didáticos. Do ponto de vista computacional, o banco representa um repositório de imagens de parasitas que poderá ser muito útil para se testar técnicas que envolvam métodos de reconhecimento de padrões. Nesse sentido, novos algoritmos poderiam ser testados usando esse banco de imagens validadas como referência (*golden standard*).

Além das imagens propriamente ditas, o banco disponibiliza também os dados morfométricos dessas imagens. Assim, embora dados morfométricos tenham sido classicamente utilizados para a descrição e classificação de parasitas do gênero *Eimeria* (Long et al., 1976), os resultados obtidos no presente trabalho permitem revisitar os trabalhos de parasitologia clássica dentro de um contexto mais moderno de microscopia óptica conjugada à tecnologia de captura de imagens digitais. A revisão dos quadros de classificação morfológica de parasitas poderá ser muito útil para profissionais de campo que necessitam identificar os parasitas. De fato, o nosso grupo publicou recentemente um novo quadro de descrição e identificação de espécies de *Eimeria* de galinha doméstica (Gruber et al., 2005), o qual foi distribuído na revista de divulgação de maior circulação no meio avícola (*Avicultura Industrial*), com uma tiragem nacional de mais de 25000 exemplares.

Com relação às imagens de oocistos de coelho, poucos são os dados disponíveis na literatura, especialmente em termos de morfometria. Assim, a criação de um repositório de imagens e dados morfométricos de todas as onze espécies de *Eimeria* que infectam esse hospedeiro é inédita, e poderá auxiliar enormemente os profissionais veterinários que trabalham com essa espécie animal.

7.1.8 Congruência entre distância morfológica e dados de filogenia molecular

Um aspecto onde a caracterização de formas pode ter um interessante impacto é na análise filogenética. Filogenistas clássicos costumavam empregar dados morfométricos em suas análises. Uma vez que a geração de seqüências de DNA se tornou rotineira e acessível a um grande número de pes-

quisadores, os dados morfométricos passaram a ser substituídos por dados de seqüências de DNA ou proteínas para as análises filogenéticas. Uma das grandes vantagens em se utilizar seqüências de DNA e proteínas está no fato de que cada base ou aminoácido, respectivamente, podem ser tratados como um caractere independente (com algumas exceções), e o número total de caracteres analisados é geralmente muito maior do que o que se consegue com dados morfológicos. De fato, concatenando-se seqüências de diferentes genes ou proteínas, é possível se aumentar ainda mais o número total de caracteres para a análise. Com isso, é possível se gerar análises com suportes estatísticos muito maiores do que com dados morfométricos.

Outro aspecto que facilita os estudos filogenéticos com dados moleculares é o fato de que o número de caracteres e seus estados é definido, sendo quatro para as seqüências nucleotídicas e 20 para as seqüências protéicas. No caso de dados morfométricos, os valores são discretos apenas quando relacionados com fenótipos qualitativos, como ausência ou presença de alguma estrutura anatômica, curto ou longo, liso ou rugoso, etc. Dados quantitativos podem, por sua vez ser contínuos ou descontínuos. Em grande parte das vezes, os dados obtidos são contínuos, o que significa que devem ser discretizados para poderem compor matrizes de dados aptas a serem aplicadas em métodos filogenéticos. Uma das críticas em relação ao uso de dados contínuos, segundo alguns autores, é que não deveria existir sobreposição de valores entre diferentes taxa, isto é, a faixa de variação de um trato em um dado táxon, não deveria conter valores que se localizam dentro da faixa de variação de um outro táxon (Thiele, 1993).

Os critérios para a discretização de dados morfométricos, por sua vez, são alvo de um grande número de discussões e controvérsias na literatura científica (Wiens, 2001). Afinal, quando discretizamos os dados apenas usando-se intervalos de mesmo tamanho, freqüência ou entropia, estamos na verdade utilizando critérios arbitrários ou estatísticos que muito possivelmente não têm nenhuma relação com homologia, ou seja, com ancestralidade comum (Swiderski et al., 1998). Assim, o compartilhamento de uma estrutura anatômica entre duas espécies pode ser mais facilmente relacionada a uma ancestralidade comum. Por outro lado, o fato de estas mesmas espécies compartilharem um estado de caractere, isto é, um dado intervalo de discretização, muito freqüentemente não implica que isso seja devido a qualquer evento evolutivo comum a ambas as espécies.

O problema de como discretizar dados contínuos é denominado de codificação dos dados morfológicos. Há várias propostas na literatura de como se proceder à codificação de caracteres, visando obter-se um critério relacionado com eventos evolutivos (Stevens, 1991; Archie, 1985; Wilkinson, 1995; Kornet, 1999). Contudo, a metodologia a ser utilizada está muito longe de se tornar um protocolo padrão.

Em função desse aspecto, decidimos realizar uma discretização baseada arbitrariamente em in-

tervalos de mesma frequência, sem a pretensão inicial de deduzir eventos evolutivos a partir dos resultados. Nesse sentido, geramos uma matriz de caracteres com os devidos estados de cada espécie e usamos métodos de distância para construir uma árvore. Abordagem similar foi realizada com distintos métodos de discretização, usando-se ainda diferentes números de intervalos, mas os resultados obtidos foram em princípio muito similares entre si.

A inferência filogenética do gênero *Eimeria* foi reportada na literatura usando-se a subunidade pequena 18S do gene de rRNA (Barta et al., 1997). Nosso grupo também realizou uma inferência filogenética recentemente, usando o genoma mitocondrial completo das sete espécies de *Eimeria* de galinha doméstica (Romano, 2004). Assim, decidimos comparar a árvore de distância obtida a partir dos dados morfométricos com as árvores determinadas por métodos de filogenia molecular.

Conforme pode ser visto na Figura 6.11, a despeito de não se poder, com os atuais dados, se associar os caracteres morfológicos com eventos de evolução, o que se vê é uma boa correlação na topologia das árvores de distância baseada em morfometria e de filogenia molecular. Isso sugere que, possivelmente, exista um certo grau de correlação entre as características morfológicas extraídas das imagens, os intervalos arbitrariamente criados, e as relações evolutivas entre as espécies de *Eimeria*.

Comparando-se os resultados de Barta et al. (1997) e Romano (2004), observa-se que *E. tenella* e *E. necatrix* estão no mesmo clado, mais distantes das outras cinco espécies. *E. maxima* e *E. brunetti* também formam um clado, enquanto que *E. acervulina* e *E. mitis* formam um terceiro clado. Finalmente, *E. praecox* aparece mais próxima do clado de *E. acervulina* e *E. mitis*, ou do clado de *E. maxima* e *E. brunetti*, dependendo do método e do marcador molecular empregado. Na árvore de distância obtida através dos dados morfométricos, também se observa a constituição dos clados de *E. maxima* e *E. brunetti*, e de *E. acervulina* e *E. mitis*. *E. tenella* e *E. necatrix*, embora não formem um clado, estão próximas entre si na topologia da árvore. Finalmente, *E. praecox* está mais próxima do clado de *E. acervulina* e *E. mitis*.

Este tipo de comparação entre dados morfológicos e moleculares é relativamente comum em organismos superiores, mas em microrganismos, os estudos morfológicos têm sido bastante limitados. Portanto, o presente trabalho apresenta dados interessantes e inéditos, uma vez que demonstra a congruência entre dados morfológicos e moleculares, em um conjunto de organismos unicelulares.

7.1.9 Perspectivas futuras

Segmentação automática e análise em lote (“*batch*”)

O sistema, na forma como está implementado atualmente, ainda requer alguma intervenção humana antes que o programa possa realizar a classificação do objeto consulta. Assim, é necessário se realizar o recorte manual das regiões contendo os objetos de interesse, utilizando-se editores de imagens. Além disso, uma vez tendo-se a imagem com um único oocisto, é necessário se proceder a um processo interativo de segmentação através da interface *web*, na qual o melhor valor do limiar (*threshold*) tem que ser testado manual e interativamente. Esta última característica implica, ainda, que seja atualmente impossível se submeter um conjunto de múltiplas imagens para análise.

Utilizando-se técnicas de processamento de imagens, é possível fazer com que esse procedimento de segmentação seja automatizado com um certo grau de confiabilidade (Yang et al., 2001; Koprowski e Wrobel, 2005; Dima et al., 2002; Nattkemper, 2004), o que permitiria obter diretamente o contorno do objeto que serviria como entrada para o processo de extração de características. Alternativamente, enquanto esse desenvolvimento não estiver concluído, é possível ainda se determinar os valores de *threshold* de forma empírica e iterativa, de tal maneira que, além do valor *default* atualmente empregado, possam ser testados, de forma recursiva, alguns valores inferiores e superiores e determinado qual deles resulta em sucesso na segmentação.

Uma vez em posse de um sistema que seja capaz de lidar com a etapa de segmentação sem intervenção do usuário, seria possível então submeter um arquivo comprimido, contendo múltiplas imagens. Isso certamente aumentaria muito a aplicação do programa e a confiança do usuário no resultado diagnóstico, uma vez que, ao invés de obter o resultado para uma única imagem, o usuário final obteria um diagnóstico para dezenas de imagens, diluindo assim possíveis erros devidos a variações morfológicas individuais.

Finalmente, um passo seguinte seria o de processar automaticamente uma imagem digital de um campo microscópico contendo múltiplos oocistos. Assim, o programa seria capaz de localizar automaticamente os objetos de interesse (os oocistos), segmentá-los individualmente e submeter as respectivas imagens para o sistema de reconhecimento de imagem. Isso eliminaria praticamente qualquer intervenção manual do usuário, além da submissão propriamente dita da imagem digital do campo microscópico.

Aprendizado online

O sistema atual foi treinado com diferentes tipos de cepas das distintas espécies de *Eimeria*, mas isso não significa que uma nova cepa de campo não apresente alguma variação significativa na sua morfologia. Assim, caso isso seja detectado, e a cepa isolada, a mesma poderia ser utilizada para incrementar o banco de imagens de treinamento do sistema. Dessa maneira poderia se aumentar o espectro de variabilidade com que o programa é capaz de lidar. Uma versão mais sofisticada do sistema poderia ainda permitir que usuários autorizados adicionassem esses novos elementos no conjunto de treino e se auto-treinar. De qualquer forma, esse procedimento teria de ser supervisionado e controlado, para se ter a certeza da origem das imagens e evitar a introdução de dados errados que poderiam levar a classificações erradas.

Aplicação do sistema a outros parasitas

O presente trabalho permitiu demonstrar que é possível se montar um sistema integrado de diagnóstico de parasitas, utilizando-se processamento digital de imagens e uma interface de usuário via *web*. Entre as possibilidades para se aumentar o escopo do trabalho, vislumbra-se basicamente dois caminhos distintos, em ordem de complexidade:

1. Adaptação do sistema para parasitas cuja morfologia se enquadre no mesmo domínio de imagem.

A utilização de parasitas do gênero *Eimeria* como modelo de estudo neste trabalho, está ligada a dois fatores: a disponibilidade de grande número de parasitas devido à nossa linha de pesquisa na área, e a morfologia relativamente simples desses parasitas. De fato, oocistos de *Eimeria*, conforme foi apresentado ao longo desse trabalho, são estruturas arredondadas, as quais apresentam poucos formatos de curvatura. Este tipo de domínio de imagem também se aplica a outros parasitas da classe Coccidia, como os dos gêneros *Cryptosporidium* e *Isospora*, entre outros. Da mesma maneira, ovos de helmintos também apresentam morfologia similar e, portanto, poderiam ter suas características morfológicas extraídas, usando-se os mesmos parâmetros descritos em nosso trabalho. De fato, a morfologia relativamente simples desses parasitas explica em parte porque a maioria dos trabalhos relatados na literatura ter utilizado coccídias e ovos de helmintos para o diagnóstico por processamento de imagem.

A adaptação do sistema descrito neste trabalho para parasitas como os citados acima, e dentro do mesmo domínio de imagem, requereria somente a constituição de um banco de imagens para o treinamento, semelhantemente ao que foi feito para as espécies de *Eimeria* de coelho.

Possíveis aplicações, extremamente úteis, incluiriam a discriminação de ovos de helmintos de grandes e pequenos ruminantes, visto que estas espécies de hospedeiros podem albergar um número relativamente grande de diferentes espécies de parasitas, sendo que algumas delas requerem tratamentos com drogas distintas.

2. Adaptação do sistema para parasitas com morfologia muito diferente de *Eimeria* spp.

Um outro desafio extremamente interessante, mas bastante mais complexo, está relacionado com o diagnóstico da malária humana, a qual é provocada por parasitas intracelulares do gênero *Plasmodium*. De fato, duas espécies de maior importância clínica em humanos, *P. vivax* e *P. falciparum*, têm uma morfologia relativamente semelhante e, somente pessoas qualificadas e altamente treinadas, podem realizar um diagnóstico preciso baseado na discriminação morfológica. Alguns aspectos tornam o diagnóstico desses parasitas muito mais complexo do que o de *Eimeria*. Primeiramente, esses protozoários apresentam um tamanho muito menor do que o de oocistos de *Eimeria*. Enquanto esses últimos apresentam um diâmetro maior geralmente variando entre 12 a 35 μm , as formas trofozoítas de *Plasmodium*, presentes em hemáceas infectadas, geralmente não ultrapassam 1-1,5 μm . Além disso, como a maior parte dos estágios é intracelular, os parasitas têm que ser identificados no interior dessas células. Além da morfologia dos parasitas variar entre as diferentes espécies, a infecção por cada um delas também altera de forma diferente a morfologia das próprias hemáceas. Para se proceder a um processamento digital de imagem para estes parasitas, requer-se um maior aumento, o que é conseguido com o uso de uma objetiva de imersão com aumento de 100 $\times$. A seguir, é preciso se segmentar as próprias hemáceas para, em um segundo momento, se segmentar os parasitas. Isso pode ser feito mediante o uso de dois valores de *threshold*. Posteriormente, o sistema tem que identificar quais hemáceas estão infectadas e, finalmente, destas, qual a espécie de *Plasmodium* presente.

Em função de toda a complexidade envolvida, poucos ainda foram os avanços para o desenvolvimento de um sistema de diagnóstico por imagem da malária. Em um artigo recente (Ross et al., 2006), os autores trabalharam com as quatro espécies de *Plasmodium* que infectam o homem (*P. falciparum*, *P. vivax*, *P. malariae* e *P. ovale*) e, segundo seus cálculos, 10 entre cada 11 hemáceas infectadas foi detectada como tal. Os resultados do segundo classificador foram, entretanto, bastante inferiores, com uma sensibilidade variando entre 29% a 85%, dependendo da espécie de *Plasmodium*.

A utilização de diferentes métodos de segmentação e de extração de características morfológicas de parasitas do gênero *Plasmodium*, poderá resultar, no futuro, em um sistema mais confiável e aplicável no campo, com enormes benefícios para o diagnóstico e escolha do trata-

mento de pacientes, especialmente em regiões com menos acesso a profissionais qualificados.

7.2 Conclusões

Nesta tese, foi proposta uma abordagem efetiva para a caracterização automática de forma de *Eimeria* spp. para a diferenciação de espécies. As características extraídas identificam as distintas propriedades dos oocistos referentes à caracterização da forma, geometria e estrutura interna. Esta representação da forma foi testada na diferenciação das sete espécies de *Eimeria* da galinha doméstica e os resultados obtidos mostraram a robustez do conjunto de características. Adicionalmente, foram aplicadas duas técnicas de discriminação, baseadas em classificadores Bayesianos, onde a primeira, que apresenta resultados por similaridade, consegue uma melhor taxa de acerto; enquanto a segunda, que apresenta resultados probabilísticos, mostra-se com melhor desempenho. Foi desenvolvido um sistema integrado de diagnóstico em tempo real, utilizando-se uma interface *web*. Além disso, foi criado um repositório público de imagens dos parasitas. A metodologia proposta também foi testada com as onze espécies de *Eimeria* que infectam o coelho doméstico, e os resultados obtidos foram similares em acerto. Finalmente, foram apresentados resultados inéditos de distância morfológica entre as diferentes espécies de *Eimeria* de galinha e sua comparação com árvores filogenéticas obtidas com marcadores moleculares. Os resultados apresentados revelaram uma grande concordância entre os resultados morfológicos e moleculares.

Referências Bibliográficas

- Adoutte, A., Tillier, S., e DeSalle, R. (1998). Commemoration of the “Molecules and Morphology in Systematics” meetings held in paris, france, march 24-march 28, 1997. *Molecular Phylogenetics and Evolution*, 9(3):331–2.
- Allen, P. e Fetterer, R. (2002). Recent advances in biology and immunobiology of *Eimeria* species and in diagnosis and control of infection with these coccidian parasites of poultry. *Clinical Microbiology Reviews*, 15(1):58–65.
- Anderson, J. (1995). *An Introduction to Neural Networks*. The MIT Press, Cambridge, Mass.
- Archie, J. (1985). Methods for coding variable morphological features for numerical taxonomic analysis. *Systematic Zoology*, 34(3):326–345.
- Ashley, S. (2006). Sharp shooter. *Scientific American*, (April).
- Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.
- Baldi, P. e Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach*. MIT Press.
- Ballard, D. H. e Brown, C. M. (1982). *Computer Vision*. Englewood Cliffs, N.J.:Prentice-Hall.
- Barlow, H. (1994). *What is the computational goal of the neocortex?*. Em C. Koch e J.L. Davis (Eds.) *Large-Scale Neuronal Theories of the Brain*. The MIT Press.
- Barta, J., Coles, B., Schito, M., Fernando, M., Martin, A., e Danforth, H. (1998). Analysis of infraspecific variation among five strains of *Eimeria maxima* from North America. *International Journal of Parasitology*, 28(3):485–92.
- Barta, J., Martin, D., Liberator, P., Dashkevich, M., Anderson, J., Feighner, S., Elbrecht, A., Perkins-Barrow, A., Jenkins, M., Danforth, H., Ruff, M., e Profous-Juchelka, H. (1997). Phylogenetic

- relationships among eight *Eimeria* species infecting domestic fowl inferred using complete small subunit ribosomal DNA sequences. *Journal of Parasitology*, 83(2):262–271.
- Bellman, R. E. (1961). *Adaptive Control Processes: A Guide Tour*. Princeton, N.J. : Princeton University Press.
- Benton, M. (1999). Early origins of modern birds and mammals: Molecules vs. morphology. *Bio-Essays*, 21(12):1043–1051.
- Bookstein, F. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.
- Box, G. E. e Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Pub. Co.
- Braddick, O., Campbell, F., e Atkinson, J. (1978). *Channels in vision: Basic aspects*. Em R. Held e H. Leibowitz e H.L. Teuber (Eds.), *Handbook of Sensory Physiology*, volume VIII. Berlin:Springer.
- Bradley, A. (1997). Use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.
- Brigham, E. O. (1988). *The Fast Fourier Transform and Applications*. Englewood Cliffs, NJ: Prentice Hall.
- Bruno, O. M., Cesar Jr., R. M., Consularo, L., e Costa, L. F. (1998). Automatic feature selection for biological shape classification in Σ YNERGOS. *SIBIGRAPI – International Symposium on Computer Graphics, Image Processing and Vision*, page 363.
- Campos, T. E. (2001). *Técnicas de seleção de características com aplicações de reconhecimento de faces*. Dissertação de Mestrado, Departamento de Ciências da Computação - Instituto de Matemática e Estatística - Universidade de São Paulo.
- Carlin, B. P. e Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall.
- Castanón, C. (2003). *Recuperação de imagens por conteúdo através de análise multiresolução por Wavelets*. Dissertação de Mestrado, Departamento de Ciências da Computação - Instituto de Ciências Matemática e de Computação - Universidade de São Paulo.

- Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. *Proceedings of the European working session on learning on Machine learning table of contents*, pages 164–178.
- Cavalli-Sforza, L. e Edwards, A. (1967). Phylogenetic analysis: Models and estimation procedures. *Evolution*, 21(3):550–570.
- Cesar Jr., R. M. e Costa, L. F. (1995). Piecewise linear segmentation of digital contours in $O(N \cdot \log(N))$ through a technique based on effective digital curvature estimation. *Real-Time Imaging*, 1(6):409–417.
- Cesar Jr., R. M. e Costa, L. F. (1996). Towards effective planar shape representation with multiscale digital curvature analysis based on signal processing techniques. *Pattern Recognition*, 29(9):1559–1569.
- Cesar Jr., R. M. e Costa, L. F. (1997). Application and assessment of multiscale bending energy for morphometric characterization of neural cells. *Review of Scientific Instruments*, 68:2177–2186.
- Chapman, H., Cherry, T., Danforth, H., Richards, B., Shirley, M., e Williams, R. (2002). Sustainable coccidiosis control in poultry production: The role of live vaccines. *International Journal of Parasitology*, 32(5):617–629.
- Chou, K. e Zhang, C. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, 30(4):275–349.
- Coelho, R. C., Gesù, V. D., Bosco, G. L., Tanaka, J. S., e Valenti, C. (2002). Shape-based features for cat ganglion retinal cells classification. *Real-Time Imaging*, 8(3):213–226.
- Comaniciu, D., Meer, P., e Foran, D. J. (1999). Image-guided decision support system for pathology. *Machine Vision and Applications*, 11(4):213–224.
- Conway, D. P. e McKenzie, M. E. (1991). *Poultry Coccidiosis, Diagnostic and Testing Procedures*. Pfizer, Inc., New York, NY, 2nd ed. edition.
- Costa, L. F., Campos, A., e Manoel, E. (2001). An integrated approach to shape analysis: Results and perspectives. *International Conference on Quality Control by Artificial Vision, Le Cresout, France*, pages 23–34.
- Costa, L. F. e Cesar Jr., R. M. (2000). *Shape Analysis and Classification: Theory and Practice*. CRC Press, Inc.

- Costa, L. F., dos Reis, S. F., Arantes, R. A., Alves, A. C., e Mutinari, G. (2004). Biological shape analysis by digital curvature. *Pattern Recognition*, 37(3):515–524.
- Crammer, K. e Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2(5):265–292.
- Cristianini, N. e Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.
- Current, W., Upton, S., e Long, P. (1990). *Taxonomy and life cycles*. Em Long P.L. (Ed.) *Coccidiosis of Man and Domestic Animals*, chapter 1, pages 1–16. Boca Raton, FL, CRC Press, Inc.
- D'Arcy Thompson, W. (1942). *On Growth and Form*. London: Cambridge.
- Darwin, C. (1859). *On the Origin of Species*. John Murray, London.
- Daugman, J. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179.
- Dauguschies, A., Imarom, S., e Bollwahn, W. (1999). Differentiation of porcine *Eimeria* spp. by morphologic algorithms. *Veterinary Parasitology*, 81(1):201–210.
- Diebolt, J. e Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375.
- Dima, A., Scholz, M., e Obermayer, K. (2002). Automatic segmentation and skeletonization of neurons from confocal microscopy images based on the 3-D wavelet transform. *Image Processing, IEEE Transactions on*, 11(7):790–801.
- Dougherty, J., Kohavi, R., e Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Proceedings of the Twelfth International Conference on Machine Learning*, 202.
- Drew, M. S., Wei, J., e Li, Z.-N. (1998). Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images. *Computer Vision, 1998. Sixth International Conference on*, pages 533–540.
- Dryden, I. e Mardia, K. (1998). *Statistical Shape Analysis*. Wiley New York.
- Duda, R. O., Hart, P. E., e Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons.

- Eck, R. e Dayhoff, M. (1966). Atlas of protein sequence and structure. *National Biomedical Research Foundation, Silver Spring, Md.*
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):882–891.
- Fayyad, U. e Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 2.
- Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22(3):240–249.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates Sunderland, Mass., USA.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.
- Fernandez, S., Costa, A., Katsuyama, A., Madeira, A., e Gruber, A. (2003a). A survey of the inter and intraspecific RAPD markers of *Eimeria* spp. of the domestic fowl and the development of reliable diagnostic tools. *Parasitology Research*, 89(6):437–445.
- Fernandez, S., Pagotto, A., Furtado, M., Katsuyama, A., Madeira, A., and Gruber, A. (2003b). A multiplex PCR assay for the simultaneous detection and discrimination of the seven *Eimeria* species that infect domestic fowl. *Parasitology*, 127(4):317–325.
- Fitch, W. (1977). On the problem of discovering the most parsimonious tree. *The American Naturalist*, 111(978):223–257.
- Flach, P. (2004). The many faces of ROC analysis in machine learning. *Tutorial in ICML2004*.
- Freeman, S. e Herron, J. C. (2004). *Evolutionary Analysis*. Upper Saddle River, NJ : Pearson/Prentice Hall.
- Gelfand, A. e Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.
- Gelfand, A., Dey, D., e Chang, H. (1992). *Model determination using predictive distributions with implementation via sampling-based methods (with discussion)*. Em Bernardo J.M. e Berger J.O. e

- Dawid A.P. e Smith A.F.M. (Eds.) Bayesian Statistics 4, pages 147–167. Oxford: Oxford University Press.
- Gonzales, R. e Woods, R. (2002). *Digital Image Processing*. Prentice–Hall, Upper Saddle River, New Jersey.
- Gruber, A., Solis, J., Castanon, C., Costa, L., Fraga, J., Fernandez, S., Barreiro, H., e Gessulli Neto, O. (2005). Entendendo e controlando a coccidiose aviária. *Gessulli Agribusiness (Poster de divulgação científica)*.
- Gudivada, V. e Raghavan, V. (1995). Content based image retrieval systems. *Computer*, 28(9):18–22.
- Hall-Beyer, M. (2005). *The GLCM Tutorial Home Page*. University of Calgary, <http://www.fp.ucalgary.ca/mhallbey/tutorial.htm>, 2.8 edition. Visited in April 12, 2006.
- Haralick, R. e Shapiro, L. (1992). *Computer and Robot Vision*. Addison-Wesley, Reading, MA.
- Haralick, R. M., Shanmugam, K., e Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.
- Hastie, T., Tibshirani, R., e Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.
- Hearn, D. e Baker, M. (1997). *Computer Graphics, C Version*. Prentice Hall, Inc.
- Huber, R., Ramonser, H., Mayer, K., Penz, H., e Rubik, M. (2005). Classification of coins using an eigenspace approach. *Pattern Recognition Letters*, 26(1):61–75.
- Huxley, J. S. (1993). *Problems of Relative Growth*. Baltimore : Johns Hopkins University.
- International Telecommunication Union (1998). ITU-R Recommendation BT.470-6. Technical report.
- Jain, A. e Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Recognition Analysis and Machine Intelligence*, 19(2):153–158.
- Jain, A. K., Duin, R. P., e Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Trans. on Pattern Recognition Analysis and Machine Intelligence*, 22(1):04–37.
- Jalba, A. C., Wilkinson, M. H., e Roerdink, J. B. (2006). Shape representation and recognition through morphological curvature scale spaces. *IEEE Transactions on Image Processing*, 15(2):331–341.

- Jenner, R. (2004). Accepting partnership by submission? Morphological phylogenetics in a molecular millennium. *Systematic Biology*, 53(2):333–342.
- Joachim, A., Dulmer, N., e Dauschies, A. (1999). Differentiation of two *Oesophagostomum* spp. from pigs, *O. dentatum* and *O. quadrispinulatum*, by computer-assisted image analysis of fourth-stage larvae. *Parasitology International*, 48(1):63–71.
- Jobanputra, R. e Clausi, D. A. (2006). Preserving boundaries for image texture segmentation using grey level co-occurring probabilities. *Pattern Recognition*, 39(2):234–245.
- Julesz, B. (1995). *Dialogues on Perception*. Bradford Book.
- Kawazoe, U. (1993). Biologia. *Simpósio Internacional sobre Coccidiose*. Fundação APINCO de Ciência e Tecnologia Avícolas, pages 1–6.
- Kersten, D. (1987). Predictability and redundancy of natural images. *Optical Society of America*, 4(12):2395–2400.
- Kittler, J., Pavel, P., e Somol, P. (2001). Advances in statistical feature selection. *Proceedings of the Second International Conference on Advances in Pattern Recognition*, pages 425–434.
- Koprowski, R. e Wrobel, Z. (2005). Automatic segmentation of biological cell structures based on conditional opening e closing. *Machine Graphics & Vision International Journal*, 14(3):285–307.
- Kornet, D. (1999). Coding polymorphism for phylogeny reconstruction. *Systematic Biology*, 48(2):365–379.
- Kucera, J. e Reznicky, M. (1991). Differentiation of species of *Eimeria* from the fowl using a computerized image-analysis system. *Folia Parasitol*, 2(38):107–113.
- Lestrel, P. E. (1997). *Fourier Descriptors and their Applications in Biology*. Cambridge University Press.
- Lestrel, P. E. (2000). *Morphometrics for the Life Sciences*. World Scientific Publishing Co. Pte. Ltd.
- Levine, M. D. (1985). *Vision in Man and Machine*. McGraw-Hill Publishing Company.
- Liew, A., Yan, H., e Yang, M. (2005). Pattern recognition techniques for the emerging field of Bioinformatics: A review. *Pattern recognition*, 38(11):2055–2073.
- Long, P. e Joyner, L. (1984). Problems in the identification of species of *Eimeria*. *Journal of Protozoology*, 4(31):535–541.

- Long, P., Millard, B., Joyner, L., e Norton, C. (1976). A guide to laboratory techniques used in the study and diagnosis of avian coccidiosis. *Folia Veterinaria Latina*, 6(3):201–217.
- Long, X., Cleveland, W. L., e Yao, Y. L. (2005). Effective automatic recognition of cultured cells in bright field images using fisher's linear discrimination preprocessing. *Image and Vision Computing*, 23(1):1203–1213.
- Madigan, D. e Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.
- Marcus, L. (1996). *Advances in Morphometrics*. Plenum Press New York.
- Marr, D. (1982). *Vision*. W.H. Freeman and Company.
- McDougald, L. e Reid, W. (1997). *Coccidiosis*. Em Calnek, B.W., Barnes, H.J., Beard, C.W., McDougald, L.R., Saif, Y.M. (Ed.) Diseases of Poultry, pages 865–883. 10a Ed. Iowa State University Press, Ames, USA.
- Minkel, J. (2006). Camera reconstructs image from single pixel. *Scientific American*, (October).
- Müller, H., Michoux, N., Bandon, D., e Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23.
- Narendra, P. M. e Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computers*, 26(9):917–922.
- Nattkemper, T. (2004). Automatic segmentation of digital micrographs: A survey. *Proc. 11th World Congr. Medical Informatics (MEDINFO)*, Pt, 2:847–851.
- Nei, M. (1975). *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam.
- Nei, M. e Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, USA.
- Olshausen, B. A. e Field, D. J. (2000). Vision and the coding of natural images. *American Scientist*, 88(3):238–245.
- Patterson, C. (1987). *Molecules and Morphology in Evolution: Conflict Or Compromise?* Cambridge University Press.

- Pereira, C. (2006). *Estatística Aplicada: Uma Visão Genuinamente Bayesiana*. Botucatu: Instituto de Biociências, Universidade Estadual Paulista.
- Pereira, C. e Stern, J. (1999). Evidence and credibility: A full Bayesian test of precise hypothesis. *Entropy*, 1:99–110.
- Pereira, C. e Stern, J. (2001). Model selection: Full Bayesian approach. *Environmetrics*, 12(6):559–568.
- Pernkopf, F. (2005). Bayesian network classifier versus selective k -NN classifier. *Pattern Recognition*, 38(1):1–10.
- Plitt, A., Imarom, S., Joachim, A., e Dauschies, A. (1999). Interactive classification of porcine *Eimeria* spp. by computer-assisted image analysis. *Veterinary Parasitology*, 86(1):105–112.
- Provost, F. e Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Em *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 43-48).
- Regan, D. (2002). A hypothesis-based approach to clinical psychophysics and to the design of visual tests. *Investigative Ophthalmology and Visual Science*, 43(5):1311–1323.
- Regan, D. M. (2000). *Human Perception of Objects*. York University.
- Rohlf, F. (1999). Shape statistics: Procrustes superimpositions and tangent spaces. *Journal of classification*, 16(2):197–223.
- Rohlf, F. e Marcus, L. (1993). A revolution in morphometrics. *Trends in Ecology and Evolution*, 8(4):129–132.
- Rohlf, F. e Slice, D. (1990). Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, 39(1):40–59.
- Romano, C. M. (2004). *Caracterização molecular e análise comparativa de genomas mitocondriais de Eimeria spp. de galinha doméstica*. Dissertação de Mestrado, Biologia da Relação Patógeno-Hospedeiro, Instituto de Ciências Biomédicas - Universidade de São Paulo.
- Ross, N. E., Pritchard, C. J., Rubin, D. M., e Dusé, A. G. (2006). Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical and Biological Engineering and Computing*, 44(5):427–436.

- Sabino, D. U., Costa, L. F., Rizzatti, E. G., e Zago, M. A. (2004). A texture approach to leukocyte recognition. *Real-Time Imaging*, 10(4):205–216.
- Sahoo, P., Soltani, S., Wong, A., e Chen, Y. (1988). A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing*, 41(2):233–260.
- Salinas-Torres, V., Pereira, C., e Tiwari, R. (2002). Bayesian nonparametric estimation in a series system or a competing-risks model. *Journal of Nonparametric Statistics*, 14(4):449–458.
- Sampat, M., Bovik, A., Aggarwal, J., e Castleman, K. (2005). Supervised parametric and non-parametric classification of chromosome images. *Pattern Recognition*, 38(1):1209–1223.
- Schnitzler, B. E., Thebo, P., Mattsson, J., Tomley, F., e Shirley, M. (1998). Development of a diagnostic PCR assay for the detection and discrimination of four pathogenic *Eimeria* species of the chicken. *Avian Pathology*, 27(5):490–497.
- Schnitzler, B. E., Thebo, P., Mattsson, J., Tomley, F., e Shirley, M. (1999). PCR identification of chicken *Eimeria*: A simplified read-out. *Avian Pathology*, 28(1):89–93.
- Scotland, R., Olmstead, R., e Bennett, J. (2003). Phylogeny reconstruction: The role of morphology. *Systematic Biology*, 52(4):539–548.
- Shen, H.-B. e Chou, K.-C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22(14):1717–1722.
- Shirley, M. (1975). Enzyme variation in *Eimeria* species of the chicken. *Parasitology*, 71(3):369–376.
- Shirley, M. (1997). *Eimeria* spp. from the chicken: Occurrence, identification and genetics. *Acta Veterinaria Hungarica*, 45(3):331–347.
- Shirley, M. e Harvey, D. (1996). *Eimeria tenella*: Infection with a single sporocyst gives a clonal population. *Parasitology*, 112(6):523–528.
- Small, C. (1996). *The Statistical Theory of Shape*. Springer.
- Smeulders, A., Worring, M., Santini, S., Gupta, A., e Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380.
- Sokal, R. e Michener, C. (1958). *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas.

- Sommer, C. (1998a). Quantitative characterization, classification and reconstruction of oocyst shapes of *Eimeria* species from cattle. *Parasitology*, 116(1):21–28.
- Sommer, C. (1998b). Quantitative characterization of texture used for identification of eggs of bovine parasitic nematodes. *Journal of Helminthology*, 72(2):179–182.
- Sonka, M., Hlavac, V., e Boyle, R. (1999). *Image Processing, Analysis and Machine Vision*. SPIE.
- Spackman, K. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the sixth international workshop on Machine learning table of contents*, pages 160–163.
- Stevens, P. (1991). Character states, morphological variation, and phylogenetic analysis: A review. *Systematic Botany*, 16(3):553–583.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.
- Swets, J. A., Dawes, R. M., e Monahan, J. (2000). Better decisions through science. *Scientific American*, 283:82–87.
- Swiderski, D., Zelditch, M., e Fink, W. (1998). Why morphometrics is not special: Coding quantitative data for phylogenetic analysis. *Systematic Biology*, 47(3):508–519.
- Swiderski, D., Zelditch, M., e Fink, W. (2002). Comparability, morphometrics and phylogenetic systematics. *Morphology, shape and phylogeny*.(MacLeod, N., e PL Forey, eds.), pages 67–99.
- Swofford, D., Olsen, G., Waddell, P., e Hillis, D. (1996). Phylogenetic inference. *Molecular Systematics*, 2:407–514.
- Takemura, C., Cesar-Jr, R., Arantes, R., Costa, L., Hingst-Zaher, E., Bonato, V., e Reis, S. (2004). Morphometrical data analysis using wavelets. *Real-Time Imaging*, 10(4):239–250.
- Theodoridis, S. e Koutroumbas, K. (1998). *Pattern Recognition*. Academic Press.
- Thiele, K. (1993). The holy grail of the perfect character: The cladistic treatment of morphometric data. *Cladistics*, 9:275–304.
- Trattner, S., Greenspan, H., Tepper, G., e Abboud, S. (2004). Automatic identification of bacterial types using statistical imaging methods. *IEEE Transaction on Medical Imaging*, 23(7):807–820.
- Tuceryan, M. e Jain, A. K. (1998). *The Handbook of Pattern Recognition and Computer Vision*, chapter Texture Analysis, pages 207–247. World Scientific Publishing Co., 2nd edition edition.

- Valafar, F. (2002). Pattern recognition techniques in microarray data analysis: A survey. *Special issue of Annals of New York Academy of Sciences, Techniques in Bioinformatics and Medical Informatics*, 980:41–64.
- Valiant, L. (1984). A theory of the learnable. *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445.
- Veltkamp, R. e Tanase, M. (2000). Content-based image retrieval systems: A survey. *Rapport Technique*.
- Welsh, J. e McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, 18(24):7213–7218.
- West, M. (1992). *Modeling with mixtures (with discussion)*. Em Bernardo J.M. and Berger J.O. e Dawid A.P. e Smith A.F.M. (Eds.) *Bayesian Statistics 4*, pages 503–524. Oxford: Oxford University Press.
- Weyl, H. (1980). *Symetry*. Princenton University Press.
- Widmer, K. W., Oshima, K. H., e Pillai, S. D. (2002). Identification of *Cryptosporidium parvum* oocysts by an artificial neural network approach. *Applied and Environmental Microbiology*, 68(3):1115–1121.
- Wiens, J. (2000). Coding morphological variation within species and higher taxa for phylogenetic analysis. *Phylogenetic analysis of morphological data*, pages 115–145.
- Wiens, J. (2001). Character analysis in morphological phylogenetics: Problems and solutions. *Systematic Biology*, 50(5):689–699.
- Wiens, J. (2004). The role of morphological data in phylogeny reconstruction. *Systematic Biology*, 53(4):653–661.
- Wilkinson, M. (1995). A comparison of two methods of character construction. *Cladistics*, 11(3):297–308.
- Williams, R. (1998). Epidemiological aspects of the use of live anticoccidial vaccines for chickens. *International Journal of Parasitology*, 28(7):1089–1098.
- Williams, R. (2002). Anticoccidial vaccines for broiler chickens: Pathways to success. *Avian Pathology*, 31(4):317–353.

- Williams, R., Bushell, A., Reperant, J., Doy, T., Morgan, J., Shirley, M., Yvoré, P., Carr, M., e Frémont, Y. (1996). A survey of *Eimeria* species in commercially-reared chickens in France during 1994. *Avian Pathology*, 25(1):113–130.
- Wyszecki, G. e Stiles, W. (2000). *Color Science*. New York: John Wiley.
- Xu, Y. e Uberbacher, E. C. (1996). Gene prediction by pattern recognition and homology search. *Proceeding of International Conference on Intelligent Systems for Molecular Biology*, 4:241–51.
- Yang, Y. S., Park, D. K., Kim, H. C., Choi, M. H., e Chai, J.-Y. (2001). Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network. *IEEE Transaction on Biomedical Engineering*, 48(6):718–730.
- Young, I. T., Walker, J. E., e Bowie, J. E. (2004). An analysis technique for biological shape I. *Information and Control*, 25(1):01–19.
- Zhang, D. e Lu, G. (1974). Review of shape representation and description techniques. *Pattern Recognition*, 25(4):357–370.
- Zou, K. H. (2002). Receiver operating characteristics (ROC) literature research. *Bibliografia on-line*, disponível em : <http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html>.
- Zuckerman, S. (1950). The pattern of change in size and shape. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 137(889):433–443.