UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
CURSO DE GRADUAÇÃO EM QUÍMICA INDUSTRIAL

VALBER ELIAS DE ALMEIDA

CLASSIFICAÇÃO DE BACTÉRIAS UTILIZANDO IMAGENS DIGITAIS E SPA-LDA

CAMPINA GRANDE – PB
2014

VALBER ELIAS DE ALMEIDA

CLASSIFICAÇÃO DE BACTÉRIAS UTILIZANDO IMAGENS DIGITAIS E SPA-LDA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Química Industrial da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do grau de Bacharel em Química Industrial.

Orientador: Prof. Dr. Paulo Henrique Gonçalves Dias Diniz

CAMPINA GRANDE – PB
2014

A447c Almeida, Valber Elias de.
Classificação de bactérias utilizando imagens digitais e SPA-LDA [manuscrito] / Valber Elias de Almeida. - 2014.
47 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Química Industrial) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Paulo Henrique Gonçalves Dias Diniz, Departamento de Química".

1. Bactérias. 2. Imagens digitais. 3. Algoritmo das Projeções Sucessivas. I. Título.

21. ed. CDD 616.014

VALBER ELIAS DE ALMEIDA

CLASSIFICAÇÃO DE BACTÉRIAS UTILIZANDO IMAGENS DIGITAIS E SPA-LDA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Química Industrial da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção do grau de Bacharel em Química Industrial.

Aprovada em 12 de Novembro de 2014.

Paulo Henrique Gonçalves Dias Diniz

Dr. Paulo Henrique Gonçalves Dias Diniz / UEPB
Orientador

Wanda Izabel M. de L. Marsiglia

Profª. Msc. Wanda Izabel M. Lima Marsiglia / UEPB
Examinadora

Eliane Rolim Florentino

Profª. Drª. Eliane Rolim Florentino / UEPB
Examinadora

Este TCC é dedicado a minha família, em especial aos meus Pais (Valdir e Edjane) pela dedicação com que me criaram mostrando-me os verdadeiros valores da vida, segundo os caminhos de Deus, e aos meus avós (Solon e Alzira) que me deixaram e não acompanharão daqui de perto esta grande conquista.

AGRADECIMENTOS

Ao iniciarmos nossa jornada recebendo o dom da vida, iniciamos Intrinsecamente a busca incessante pela realização de nossos objetivos, vivemos cada dia esperando um outro, onde nele estes objetivos serão alcançados. Durante essa jornada nos damos conta de que não conseguimos alcançá-los sem o apoio e ajuda de outras pessoas, que caminham em busca de objetivos comuns ou não, os que comungam dos mesmos ideais lhe seguem e estão sempre ao seu lado lhe dando apoio e lhe auxiliando a superar as barreiras, outros seguem em outras direções, cruzando por muitas vezes seus caminhos deixando lições e aprendizados. Acima de todos existe um Deus que providencia e indica caminhos possíveis para que estas pessoas encontrem-se, e sigam seus planos segundo os seus desígnios, dês de o dia em que as criou.

RM 12:3-7

...Porque pela graça, que me é dada, digo a cada um dentre vós que não saiba mais do que convém saber, mais que saiba com temperança, conforme a medida da fé que Deus repartiu a cada um. Porque assim como em um corpo temos muitos membros, e nem todos os membros têm a mesma operação. Assim nós, que somos muitos, somos um só corpo em cristo, mas individualmente somos membros uns dos outros. De modo que tendo diferentes dons, segundo a graça que nos é dada, se é profecia, seja ela segundo a medida da fé, se é ministério, seja em ministrar; se é ensinar haja dedicação ao ensino...

Agradeço a Deus primeiramente pela permissão de ser o que fui até hoje e serei até onde for os meus dias, assim como tudo que conquistei e conquistarei até lá, também a todos os que me acompanharam nesta caminhada, sem vocês nada disso seria possível. Em especial a toda minha família destacando-se meus Pais (Valdir e Edjane), Irmãos (Wiliam e Ewelin) meus avós, Tios e Primos, por serem meus pilares de sustentação; A minha melhor Amiga, Companheira e Namorada (Denise); Ao Prof Germano, meu orientador e amigo durante toda minha graduação, a quem tenho grande respeito, gratidão e admiração, por tudo que sempre fez por mim da melhor forma possível e a professora Ana Claudia pelas coorientações e concelhos; A Paulo, meu orientador no TCC, pela dedicação e amizade criada nesse período; Aos melhores amigos/irmãos "das antigas" (David, Gean, Marcelo, Clediano, Adriano, Priscila) que sempre foram fundamentais pelo que sou hoje; A os novos melhores amigos (Thomas, Martina, Jéssica, Ingredy, Ellen, Gizelly, Rossana, Juliana) assim como todos os que conviveram comigo durante todo o período de graduação tanto do LQAQ quanto do LABDEM; A todos os professores que me instruíram e me mostraram os caminhos a seguir durante toda uma vida acadêmica e a meus colegas de todas as turmas que passei.

AGRADEÇO A DEUS PELOS MEUS AMIGOS E FAMILIA, E A TODOS VOCÊS POR EXISTIREM! MUITO OBRIGADO!

SUMÁRIO

LISTA DE ABREVIATURAS

ATCC	American Type Culture Collection
CCD	Dispositivo de Carga Acoplada
DA	Análise Discriminante
DF	Função Discriminante
HSB	Matiz-Saturação-Brilho
HSI	Matiz-Saturação-Intensidade
KS	Kennard Stone
LDA	Análise Discriminante Linear
PC	Componente Principal
PCA	Análise por Componentes Principais
PLS	Mínimos Quadrados Parciais
QSAR	Relação Quantitativa Estrutura-Atividade
RGB	Vermelho-Verde-Azul
SIMCA	Modelagem Independente e Flexível por Analogia de Classes
SPA	Algoritmo das Projeções Sucessivas
UFC	Unidades Formadoras de Colônia
UV	Ultra Violeta
VC	Validação Cruzada

Resumo

Neste trabalho é proposta uma nova metodologia para a diferenciação de cinco diferentes tipos de bactérias (*Escherichia coli*, *Enterococcus faecalis*, *Streptococcus salivarius*, *Streptococcus oralis* e *Staphylococcus aureus*) utilizando histogramas de cor obtidos a partir de imagens digitais capturadas com uma *webcam* e SPA-LDA. Histogramas de cor nos canais de RGB, HSI e escala de cinza, além de suas combinações (tons de cinza + RGB; tons de cinza + HSI; tons de cinza + RGB + HSI) foram então utilizados como informação analítica. Para fins de comparação com os resultados obtidos por SPA-LDA foram empregados diferentes classificadores multivariados: SIMCA, PCA-LDA e PLS-DA. O melhor resultado de classificação foi obtido usando RGB e SPA-LDA, alcançando 94 e 100% de precisão da classificação para os conjuntos de treinamento e teste, respectivamente. Este resultado é extremamente positivo do ponto vista de análises microbiológicas e clínicas, porque evita a identificação de bactérias com base na identificação fenotípica do organismo causador usando coloração de Gram, a cultura das cepas e os testes bioquímicos. Portanto, o método proposto apresenta vantagens inerentes, promovendo uma alternativa mais simples, rápida e de baixo custo para a identificação de bactérias.

Palavras chave: bactérias; imagens digitais; Algoritmo das Projeções Sucessivas.

Abstract

In this work, a new approach is proposed to verify the differentiating characteristics of five bacteria (Escherichia coli, Enterococcus faecalis, Streptococcus salivarius, Streptococcus oralis, and Staphylococcus aureus) by using digital images obtained with a simple webcam and variable selection by the Successive Projections Algorithm associated with Linear Discriminant Analysis (SPA-LDA). In this sense, color histograms in the red–green–blue (RGB), huesaturation - value (HSV), and grayscale channels and their combinations were used as input data, and statistically evaluated by using different multivariate classifiers (Soft Independent Modeling by Class Analogy (SIMCA), Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA), Partial Least Squares Discriminant Analysis (PLS-DA) and Successive Projections Algorithm-Linear Discriminant Analysis (SPA-LDA)). The bacteria strains were cultivated in a nutritive blood agar base layer for 24 h by following the Brazilian Pharmacopoeia,maintaining the status of cell growth and the nature of nutrient solutions under the same conditions. The best result in classification was obtained by using RGB and SPA-LDA, which reached 94 and 100% of classification accuracy in the training and test sets, respectively. This result is extremely positive from the viewpoint of routine clinical analyses, because it avoids bacterial identification based on phenotypic identification of the causative organism using Gram staining, culture, and biochemical proofs. Therefore, the proposed method presents inherent advantages, promoting a simpler, faster, and low-cost alternative for bacterial identification.

Keywords: Bacteria, Digital Images, Successive projections algorithm

1. INTRODUÇÃO

1.1. Caracterização da problemática

As bactérias patogênicas são as principais causadoras de uma série de doenças contagiosas responsáveis pela morte de inúmeras pessoas e animais no mundo, muitas delas causadas pela lentidão do diagnóstico ou, em alguns casos, pela falta dele [1,2]. O rápido diagnóstico destas doenças com os seus respectivos precursores é um passo primordial para o início imediato do tratamento adequado, aumentando a sua eficácia, especialmente em pacientes com sistemas imunitários enfraquecidos [3]. As metodologias de rotina empregadas nos laboratórios de microbiologia clínica para identificação destas bactérias baseiam-se nas características fenotípicas, nas propriedades bioquímicas e metabólicas do microrganismo [4]. Neste sentido, devem-se avaliar os padrões de crescimento das colônias das bactérias em meios de cultura, a morfologia das colônias, a coloração de Gram e outros testes bioquímicos. No entanto, estes métodos requerem abrangente conhecimento sobre as características individuais de cada espécie de microrganismo, além de ser muito trabalhoso e demorado, exigindo mais de 48 h para o crescimento das bactérias. Além disso, algumas cepas apresentam características bioquímicas únicas que não se encaixem em nenhum padrão usado como uma característica de qualquer gênero ou espécie conhecida [5,6].

A fim de contornar estes inconvenientes, algumas metodologias para identificação e classificação de bactérias têm sido relatadas na literatura, dentre elas cromatografia gasosa [7,8], eletroforese capilar [9], espectrometria de massas [6,10], espectroscopia Raman [11], espectroscopia no infravermelho [12–14], fluorescência [15,16] e ressonância magnética nuclear [17]. Apesar de estas técnicas analíticas instrumentais obterem resultados confiáveis e precisos na identificação e classificação de microrganismos, o uso de imagens

digitais apresenta-se como uma alternativa viável com algumas vantagens intrínsecas, uma vez que requer uma instrumentação de baixo custo, evita a manipulação da placa de Petri onde os microrganismos são cultivados e não exige o conhecimento de um microbiologista experiente para sua utilização. Por este motivo, alguns trabalhos para identificação e/ou classificação de microrganismos usando análise por imagens digitais foram propostos na literatura [18–21].

Dubuisson e colaboradores [18] desenvolveram uma metodologia utilizando segmentação de imagens digitais para a classificação de *Methanospirilium hungatei* e *Methanosarcina mazei* com base em suas formas. As culturas de *M. hungatei* foram corretamente identificadas, enquanto que os resultados para as culturas de *M. mazei* não foram tão precisos.

Kumar e Mittal [19] obtiveram a geometria e parâmetros ópticos utilizando microscopia de fluorescência e imagens para a identificação de cinco microrganismos. As imagens das cepas de *Bacillus thuringiensis*, *Escherichia coli K12*, *Lactobacillus brevis*, *Listeria innocua* e *Staphylococcus epidermidis* tingidas com dois corantes fluorescentes foram captadas com uma câmara CCD acoplada a um microscópio de luz. A emissão de fluorescência (intensidade de nível de cinza) para a *B. thuringiensis* foi a mais elevada em comparação com os outros microrganismos, enquanto a emissão para a *L. brevis* foi a mais baixa. Os valores médios de 10 percentuais de histogramas de imagens de *L. innocua* e *S. epidermidis* foram significativamente diferentes dos de *L. brevis*. Usando 99 valores percentuais, *B. thuringiensis* podem ser diferenciadas dos microorganismos restantes, assim como *E. coli* também pode ser diferenciada de *L. brevis* e *S. epidermidis*.

Huff e colaboradores [20] utilizaram um sensor de dispersão de luz para a identificação em tempo real de colônias de *Vibrio parahaemolyticus*, *Vibrio vulnificus* e *Vibrio cholerae* sobre placas de ágar sólido. As colônias foram iluminadas por um feixe de

laser em 635 nm e as assinaturas das imagens dispersadas foram adquiridas utilizando uma câmara CCD. Uma técnica de reconhecimento de padrões obteve 99% de classificação correta. A metodologia proposta detectou com sucesso *V. cholerae*, *V. parahaemolyticus* e *V. vulnificus* em amostras de ostras ou de água em 18 h, mesmo na presença de outros vibrios ou bactérias.

Suchwalko e colaboradores [21] identificaram espécies de bactérias (*Salmonella enteritidis*, *Staphylococcus aureus*, *Staphylococcus intermedius*, *Escherichia coli*, *Proteus mirabilis*, *Pseudomonas aeruginosa* e *Citrobacter freundii*) com base em padrões de difração de Fresnel das colônias registradas em um sistema óptico com iluminação por ondas esféricas convergentes. O método proposto utilizou processamento de imagens e análise estatística com base na extração e seleção de características e métodos de classificação, chegando a identificação de 98% das bactérias em 36 horas a partir da aquisição da amostra.

Todas as abordagens mencionadas anteriormente utilizam parâmetros geométricos e/ou de extração de características a partir de imagens das bactérias. Por outro lado, os histogramas de cor descrevem a distribuição estatística dos pixels de uma imagem digital como uma função do componente de cor capturada e não uma característica ou um comportamento físico-químico direto [22]. Os histogramas de cor têm sido utilizados com sucesso como informação analítica para a classificação de chás [22], méis [23], óleos vegetais comestíveis [24].

Assim, neste trabalho foi proposta uma nova metodologia para a diferenciação de cinco diferentes tipos de bactérias (*Escherichia coli*, *Enterococcus faecalis*, *Streptococcus salivarius*, *Streptococcus oralis* e *Staphylococcus aureus*) utilizando histogramas de cor obtidos a partir imagens digitais capturadas com uma *webcam* e SPA-LDA [25]. Histogramas de cor nos canais de RGB, HSI e escala de cinza, além de suas combinações

(tons de cinza + RGB; tons de cinza + HSI; tons de cinza + RGB + HSI) foram então utilizados como informação analítica. Para fins de comparação com os resultados obtidos por SPA-LDA foram empregados diferentes classificadores multivariados: SIMCA, PCA-LDA e PLS-DA.

1.2. Objetivo geral

Desenvolvimento de uma metodologia simples, rápida, de baixo custo e não destrutiva baseada na utilização de imagens digitais e técnicas de reconhecimento de padrões supervisionadas para a classificação de cinco diferentes tipos de bactérias.

1.3. Objetivos específicos

- Utilizar de histogramas de cor (RGB, HSI e escala de cinzas) gerados a partir das imagens digitais capturadas com uma *webcam*, empregando-os como dados analíticos de entrada.
- Utilizar de técnicas de reconhecimento de padrões supervisionadas (SIMCA, PCA-LDA, PLS-DA e SPA-LDA) para fins de classificação multivariada.
- Classificar amostras de bactérias dos gêneros *Escherichia*, *Enterococus*, *Streptococcus* e *Shaphylococcus.*

2. REVISÃO DE LITERATURA

2.1. BACTÉRIAS

As bactérias são microrganismos unicelulares do reino Monera que podem ser encontradas isoladas ou em forma de colônias. Podem ser classificadas em três formas básicas: cocos, bacilos e espirais. Os cocos geralmente se apresentam em forma esférica, porém, em alguns casos, podem ser ovais, alongados ou achatados em algumas extremidades. Podem ser subdivididos em: (a) **diplococos**, quando aparecem aos pares; (b) **estreptococos**, quando se dividem e permanecem ligados em forma de cadeia; (c) **tétrades** quando se dividem e permanecem em quartetos; (d) **sarcinas**, que se dividem em três planos e permanecem unidos em forma de cubo, com oito bactérias; (e) **estafilococos**, que se dividem em múltiplos planos e formam agrupamentos tipo cacho; e (f) **enterococos**, que se dividem por fissão binária para formarem cadeias de bactérias [26].

Os enterococos são adaptados às áreas do corpo ricas em nutrientes, mas pobres em oxigênio, como o trato gastrintestinal, a vagina e a cavidade oral. Também são encontrados em grandes quantidades nas fezes humanas. Como são micro-organismos relativamente resistentes, eles persistem como contaminantes em ambientes hospitalares, mãos, jogos de cama e até nos gases fecais. Em décadas recentes, eles se tornaram a principal causa de infecções nosocomiais, especialmente por sua alta resistência à maioria dos antibióticos. Duas espécies, *Enterococcus faecalis* e *Enterococcus faecium*, são responsáveis pela maioria das infecções de feridas cirúrgicas e do trato urinário [26].

As bactérias do gênero estreptococos são divididas em beta-hemolíticas e não beta-hemolíticas. Esta divisão é feita verificando o seu crescimento em meio ágar-sangue e as espécies beta-hemolíticas produzem uma hemolisina que forma um halo claro de hemólise no ágar-sangue [26].

Os estafilococos são encontrados organizados em forma de cachos de uva. Dentre as bactérias deste gênero a mais importante é o *Staphylococcus aureus*, que ganha essa denominação pela pigmentação amarelada de suas colônias. Esta bactéria produz muitas toxinas que contribuem para sua patogenicidade, aumentando sua capacidade de invadir o corpo e danificar os tecidos. A infecção de feridas cirúrgicas por *S. aureus* é um problema comum em ambientes hospitalares devido a sua capacidade em desenvolver rapidamente resistência a antibióticos como a penicilina [26].

Devido a menor capacidade de divisão dos bacilos quando comparados com os cocos a sua classificação é menor, dividindo-se basicamente em diplobacilos, estreptobacilos e cocobacilos. O exemplar mais conhecido é a *Escherichia coli*, que possui forma de bacilo e é encontrada normalmente no intestino grosso dos vertebrados. Sua presença é benéfica, pois ajuda na produção de certas vitaminas e participa da digestão de alguns alimentos que sem ela não seriam digeridos. Contudo, uma espécie chamada *E. coli O157:H7* é causadora de diarreia sanguinolenta quando cresce no intestino [26].

2.2. IMAGENS DIGITAIS

Uma imagem digital é formada por uma matriz (m × n) de linhas e colunas, respectivamente, que possui um número de estruturas em cada ponto da matriz chamada de pixels. O pixel, cujo nome é derivado da aglutinação das palavras inglesas *Picture* e *Element*, é o menor elemento de uma imagem do qual pode ser retirado uma série de informações. A distribuição estatística dos pixels como uma função do componente de cor gravado compõe o histograma de cor para uma dada imagem digital. A quantidade de linhas e colunas define a resolução espacial da imagem digital. A medida que o tamanho da

matriz aumenta sua resolução também aumentará e, em contra partida, o tamanho de cada pixel será diminuído. **[27]**

A profundidade de uma imagem corresponde à quantidade de memória do computador associada a cada pixel. A forma de armazenagem utilizada pelos computadores é em forma de códigos binários denominados “bits”, onde cada um deles pode ser desligado ou ligado e então são atribuídos valores de 1 ou 0. A quantidade de bits associadas à matriz da imagem digital, indica a quantidade de informação máxima que pode ser armazenada em cada pixel. Em imagens binárias (preto e branco), a quantidade de bits define quantos tons de cinza são apresentados. Uma imagem de 10 bits contém 1.024 tons de cinza, enquanto que uma imagem com 12 bits contém 4.096 tons de cinza [27].

Uma imagem em preto e branco ou em tons de cinza é considerada como sendo de um único canal, diferentemente das imagens coloridas que podem possuir mais que um e, em uma combinação de canais, fornecem uma grande variedade de cores. Por exemplo, uma imagem RGB contém três canais. Cada canal terá uma tonalidade de cada uma das cores, assim uma imagem RGB necessitaria de três vezes mais a quantidade de memória do computador do que a imagem armazenada em escala de cinza. Uma imagem padrão RGB é de 24 bits, sendo 8 bits para cada canal. A imagem é constituída pelos canais de três cores, cada cor tendo tonalidades de brilho entre 0 e 255 (**Figura 1**), se a imagem RGB é de 48 bits, cada um dos três canais tem uma escala de cor de 16 bits [27].

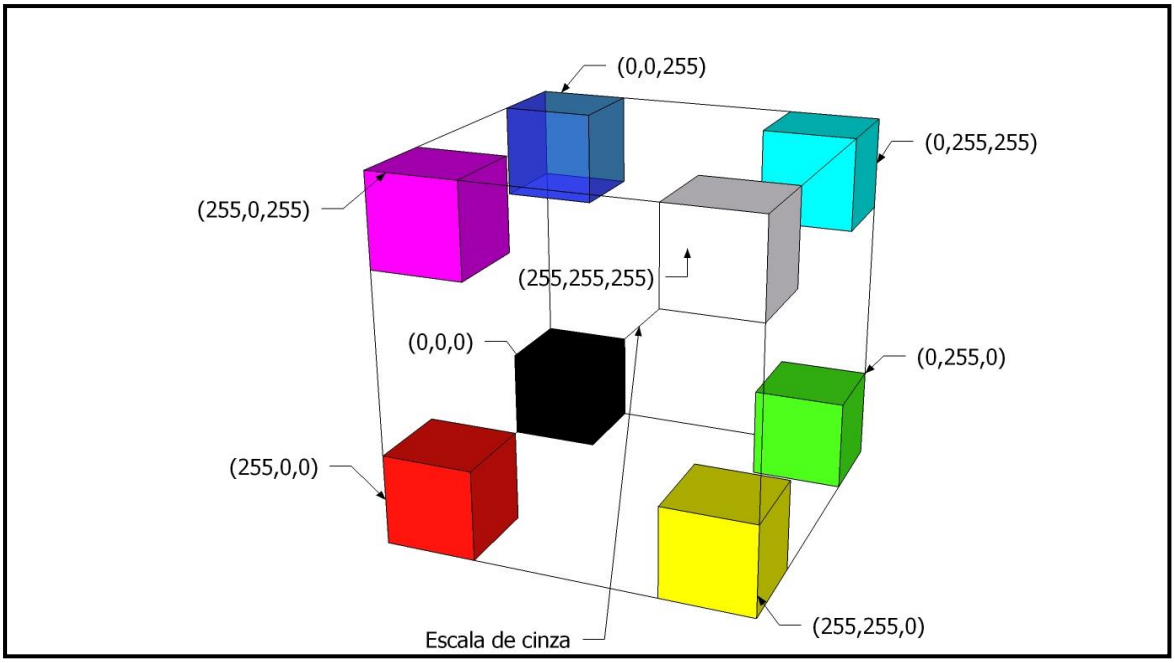

Figura 1. Modelo de cores RGB.

O sistema óptico do olho humano é capaz de distinguir centenas de milhares de tons e intensidades de cores diferentes, mais apenas cerca de 100 tons de cinza. Portanto, em uma imagem uma grande quantidade extra de informação pode estar contida na cor e esta informação adicional pode então ser usada para simplificar a análise da imagem, por exemplo, identificação de objetos e extração com base na cor. Três quantidades independentes são utilizadas para descrever qualquer cor particular, formando o sistema HSB (**Figura 2**). O matiz é determinado pelo comprimento de onda dominante. A saturação é determinada pela pureza da excitação e depende da quantidade de luz branca misturada com o matiz. O matiz puro é totalmente saturado, ou seja, nenhuma luz branca está misturada. Matiz e saturação em conjunto determinam a cromaticidade de uma determinada cor. Finalmente, intensidade é determinada pela quantidade de luz correspondendo às cores mais intensas [28].

Luz acromática não tem cor – seu único atributo é a quantidade ou intensidade. A escala de cinza é uma medida de intensidade. A intensidade é determinada pela energia e é,

por conseguinte, uma grandeza física. Por outro lado, o brilho (ou luminosidade) é determinado pela percepção da cor e é, portanto, psicológico. Dados azul e verde igualmente intensos, o azul é percebido como sendo mais escuro que o verde [29].

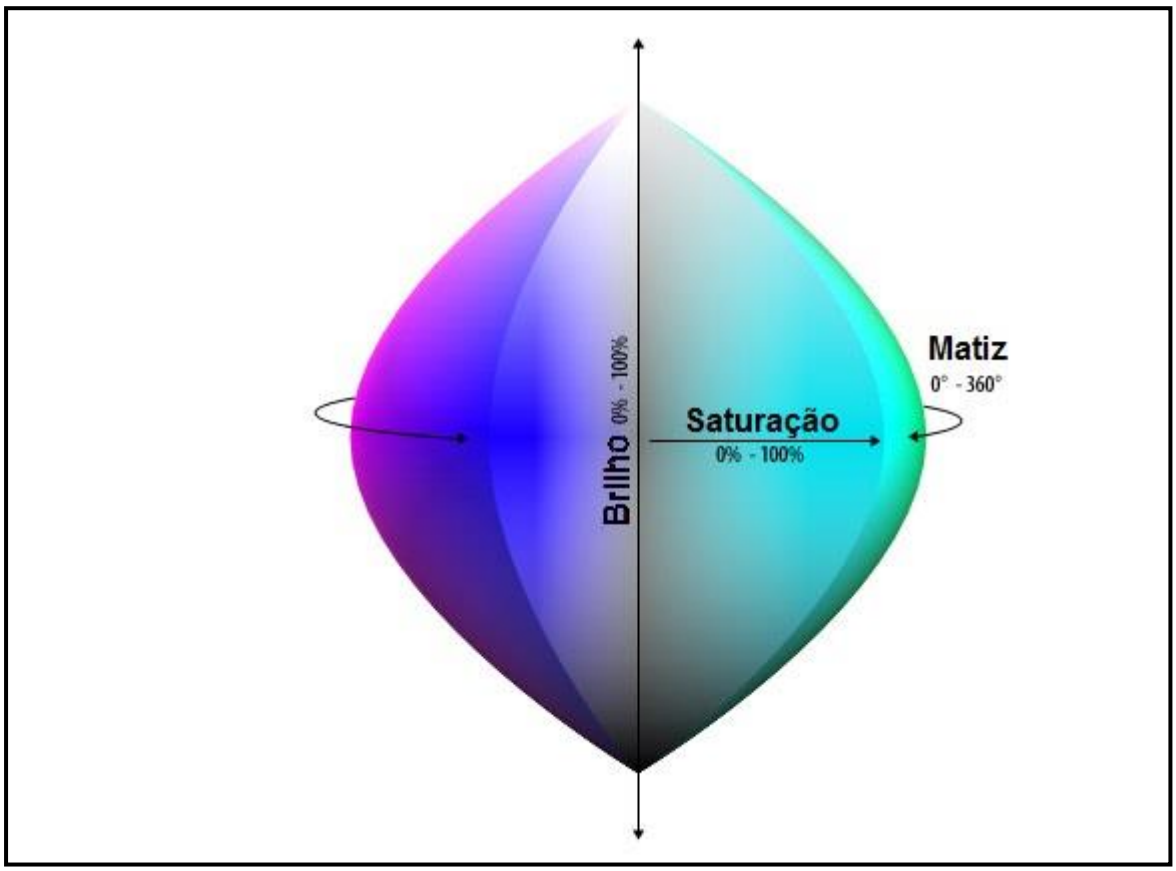

Figura 2 Sistema de cor HSB. Adaptado de http://dba.med.sc.edu/price/irf/Adobe_tg/models/hsb.html

A cor depende principalmente das propriedades de reflectância de um objeto. Vemos os raios de luz que são refletidos, enquanto outros são absorvidos. No entanto, também é preciso considerar a cor da fonte luminosa e a natureza do sistema visual humano. Por exemplo, um objeto que reflete vermelho e verde aparecerá verde quando nenhuma luz vermelha o iluminar e, inversamente, ele vai aparecer vermelho na ausência de luz verde. Em uma luz branca pura, ele aparecerá amarelo (vermelho + verde) [29].

2.3. QUIMIOMETRIA

A Quimiometria é uma área da Química Analítica que faz uso de ferramentas matemáticas e estatísticas para a resolução e análise de problemas químicos para dados multivariados. Os dados multivariados são dispostos na forma de matriz e são organizados em linhas (amostras) e colunas (variáveis) (**Figura 3**).

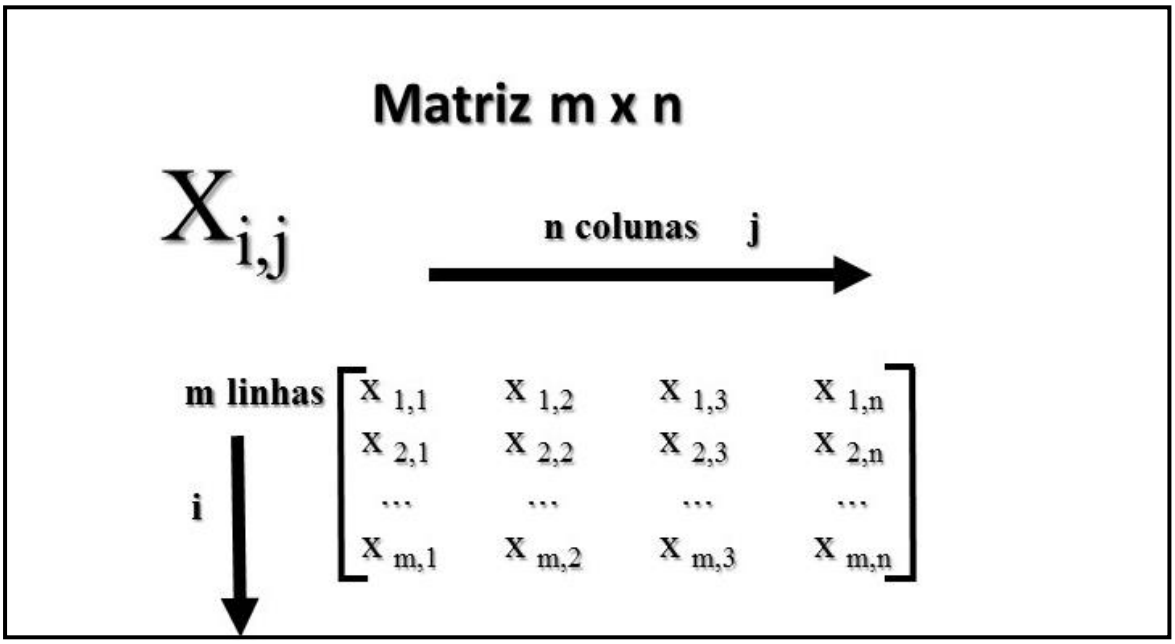

Figura 3 representação da matriz de dados

A Quimiometria é dividida basicamente em três grandes áreas: reconhecimento de padrões [25], calibração multivariada [30], planejamento e otimização de experimentos [31]. Além disso, devido à complexidade dos dados faz-se necessária a aplicação de técnicas de pré-processamento de dados, dentre as quais se destacam as técnicas de seleção de variáveis e amostras [25,32]. Uma vez que neste trabalho foram empregadas apenas as técnicas de reconhecimentos de padrões, as diferentes abordagens selecionadas são descritas a seguir. Vale ressaltar que as técnicas de reconhecimento de padrões supervisionadas utilizam a informação sobre a associação de classe das amostras para um

determinado grupo (classe ou categoria) de modo a classificar novas amostras desconhecidas em uma das classes conhecidas [33].

2.3.1. Modelagem SIMCA

A Modelagem SIMCA, é um método de reconhecimento de padrões supervisionado para classificação de amostras, utilizando a transformação de variáveis originais em variáveis latentes através da PCA. Para isto, um modelo PCA é construído para cada uma das classes de amostras; assim, a delimitação das fronteiras das classes é definida pelas PCs de cada modelo PCA e, posteriormente, uma amostra desconhecida é classificada em uma das classes, quando esta se encontra dentro das fronteiras de um dos modelos. Para a construção dos modelos delimitantes são utilizadas as amostras do conjunto de treinamento que contém amostras de todas as classes em estudo, sabendo a qual pertence cada uma destas amostras.

Na **Figura 4** encontra-se a representação gráfica da modelagem SIMCA para duas classes onde os cubos em amarelo e azul representam os modelos das classes A e B, respectivamente, onde as delimitações multidimensionais para a classe A são definidas pelas PCs 1A, 2A, e 3A, enquanto que as PCs 1B, 2B e 3B delimitam as fronteiras para a classe B.

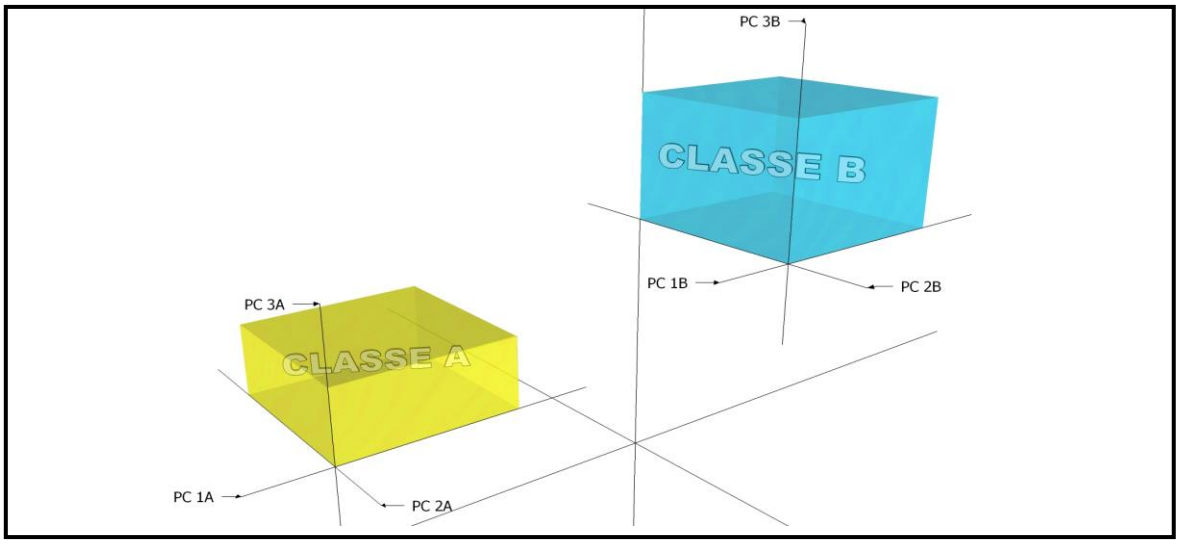

Figura 4 Representação da disposição espacial da modelagem SIMCA.

A classificação de uma amostra desconhecida D em uma classe é feita a partir do cálculo da distância desta amostra para a fronteira da classe mais próxima, segundo um grau de confiança de classificação, associando-a em uma das classes ou em nenhuma, como representado pela **Figura 5**.

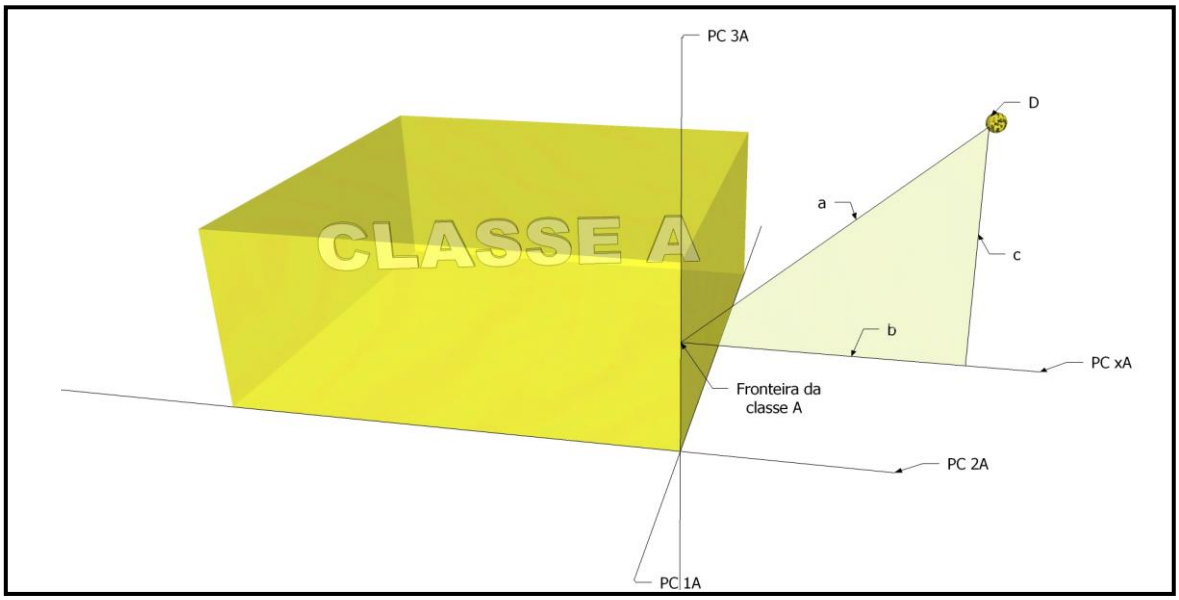

Figura 5 Representação da classificação de uma amostra desconhecida D.

A classificação de uma amostra é feita quando esta apresenta variância dentro de um valor crítico em função de "a", que é a distância entre a fronteira da classe e a amostra D a ser classificada. Esta distância pode ser estimada a partir da **Equação 1**.

$$a^2 = b^2 + c^2 \quad (1)$$

Onde "a" corresponde a distância entre a amostra D e a fronteira da classe, "b" corresponde a distância entre a fronteira da classe e a projeção da amostra D na PC xA e "c" corresponde ao resíduo da PC xA.

Após calcular-se o valor de "a", aplica-se o teste F dividindo-o pela variância da classe para a obtenção de um valor F_{cal}. Depois, escolhe-se empiricamente um F_{crit} a partir de uma tabela de teste F. Se F_{cal} for menor que F_{crit}, a amostra D é classificada como pertencente à classe em estudo [28, 34, 35].

2.3.2. Modelagem LDA

A modelagem LDA, baseia-se na determinação de funções discriminantes lineares, que maximizam a razão da variância interclasse e minimiza a razão da variância intraclasse. Uma grande limitação matemática da técnica LDA é a restrição do número de variáveis a ser estudada, que nunca poderá ser maior que o número de amostras.

2.3.2.1. PCA - LDA

Para contornar a limitação da LDA várias estratégias são utilizadas para a redução do número de variáveis, que, na maioria dos problemas em Quimiometria, devido ao avanço das técnicas instrumentais, são de ordem cada vez maior, atingindo facilmente milhares de

variáveis. A técnica de Análise de Componentes Principais reduz o número de variáveis decompondo-as em outras variáveis chamadas de variáveis latentes e, no caso da PCA, estas variáveis ganham o nome de PC, que são obtidas traçando-se vetores na direção de maior variância entre as variáveis originais como apresentado na **Figura 6a**.

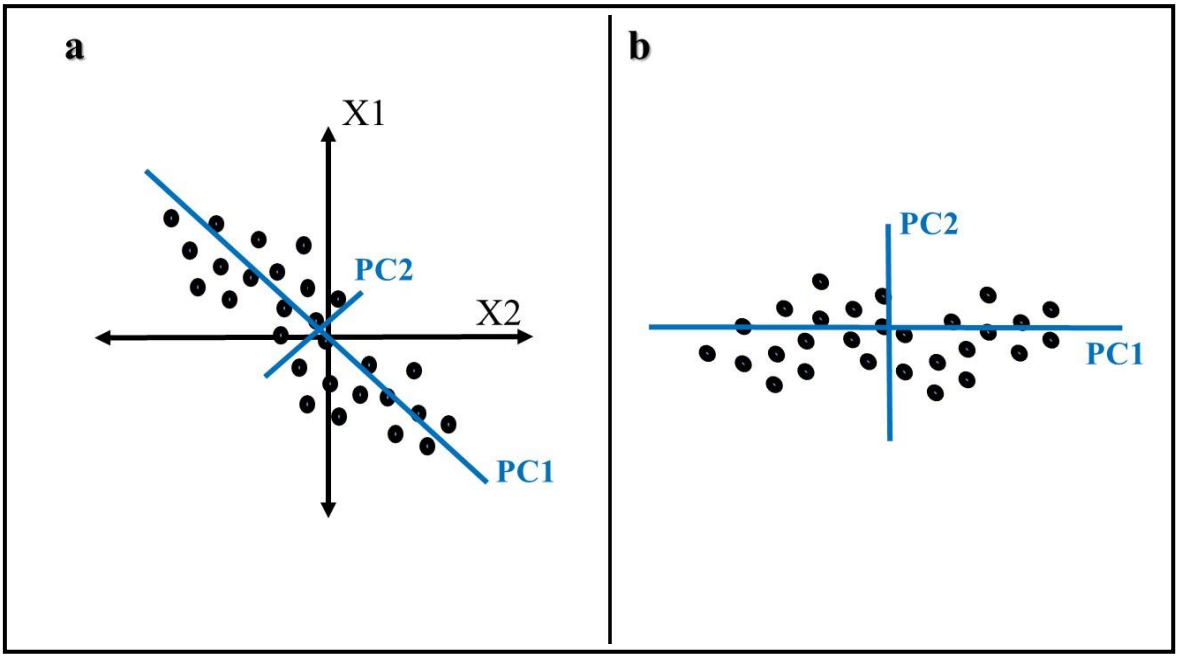

Figura 6. Decomposição das variáveis originais em componentes principais (PCs).

A segunda PC é traçada ortogonalmente à primeira, de modo a não conter na segunda PC nenhuma informação explicada pela anterior, e a PC subsequente sempre será traçada dentro do resíduo deixado pela antecendente. Estas PCs passam então a ser as novas variáveis (**Figura 6b**) e são usadas como dados de entrada para a classificação por LDA [36].

2.3.2.2. APS - LDA

O Algoritmo SPA, é uma técnica de seleção de variáveis que foi proposta inicialmente associada à Regressão Linear Múltipla para a calibração multivariada utilizando dados espectroscópicos e posteriormente adaptada para fins de classificação. O SPA é uma técnica do tipo *forward* com a restrição de que a variável incorporada em cada interação deve ser a menos colinear possível com as variáveis previamente selecionadas. As cadeias de variáveis são então sequencialmente avaliadas com base em uma função de custo; no caso de LDA é utilizada a função de custo G que indica o risco médio G de classificação incorreta pela LDA [25].

2.3.3. Modelagem PLS-DA

A modelagem PLS-DA baseia-se no algoritmo PLS2 combinado com a análise discriminante, que procura variáveis latentes com um máximo de covariância com as variáveis de categoria (Y). O novo objeto é então designado para a classe com o valor máximo do vetor Y ou, alternativamente, um limiar entre zero e um é determinado para cada classe [37].

3. MATERIAIS E MÉTODOS

3.1. Amostras

Cepas padrão ATCC dos gêneros *Escherichia, Enterococus, Streptococcus e Staphylococcus* disponibilizadas pela Fundação Oswaldo Cruz foram utilizadas neste estudo. As bactérias selecionadas foram *Escherichia coli* (14 amostras), *Enterococus faecalis* (15 amostras), *Streptococcus salivarius* (13 amostras), *Streptococcus oralis* (13 amostras) e *Staphylococcus aureus* (12 amostras).

O inóculo bacteriano foi padronizado seguindo as recomendações da Farmacopeia Brasileira [38], que se baseia na metodologia da *Clinical and Laboratory Standards* [39]. Em primeiro lugar, as amostras de bactérias foram preparados em ágar infusão de cérebro e coração (BHI), até alcançar 85% de transmitância em 630 nm num espectrofotômetro UV-Vis Biospectro, modelo SP22, a fim de se obter uma preparação bacteriana com concentração final de 10^6 UFC ml^{-1}. A seguir, 1mL desta suspensão foi diluída com 9 mL de uma solução salina de NaCl a 0,9% m v^{-1}. A camada base foi então produzida pela adição de 20 mL de um meio ágar sangue nutritivo em uma placa Petri. Após o endurecimento do ágar sangue, cinco mililitros do inóculo foram padronizados e adicionados sobre a camada base, esperando o seu re-endurecimento. As placas foram então incubadas a 37°C durante 24 horas. Esta metodologia está sumarizada na **Figura 7**.

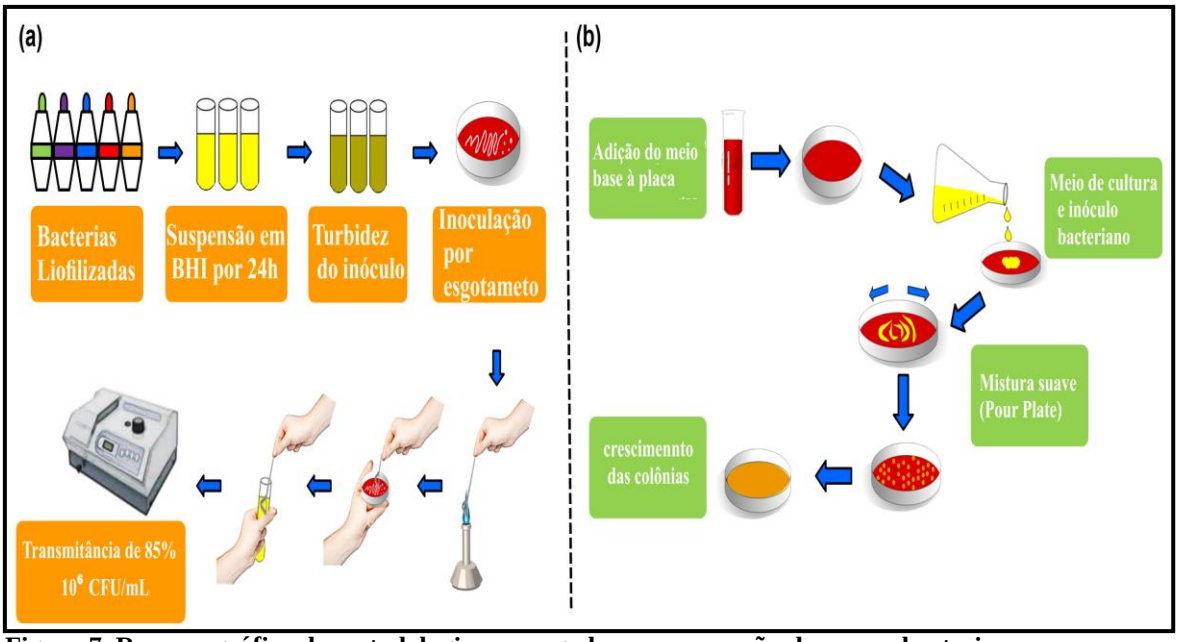

Figura 7. Resumo gráfico da metodologia empregada na preparação das cepas bacterianas.

3.2. Instrumentação

Na **Figura 8** é representado o aparato construído para aquisição de imagens das culturas de bactérias. Uma caixa com dimensões de tamanho 30 cm × 22 cm × 23 cm foi construída e revestida internamente com papel branco de escritório a fim de evitar interferências externas de luz e reflexão da luz dentro da caixa, assegurando a uniformidade para a captura das imagens [24]. Uma webcam, modelo Microsoft Lifecam Cinema, com 7.1 megapixels, foi fixada na parte central superior, posicionada no centro de uma lâmpada fluorescente circular de 8 watts para homogeneizar a iluminação interna, de modo a capturar imagens da amostra na posição vertical. A distância entre a câmara e o suporte da amostra foi de 15 cm e entre a iluminação e o suporte da amostra foi de 25 cm.

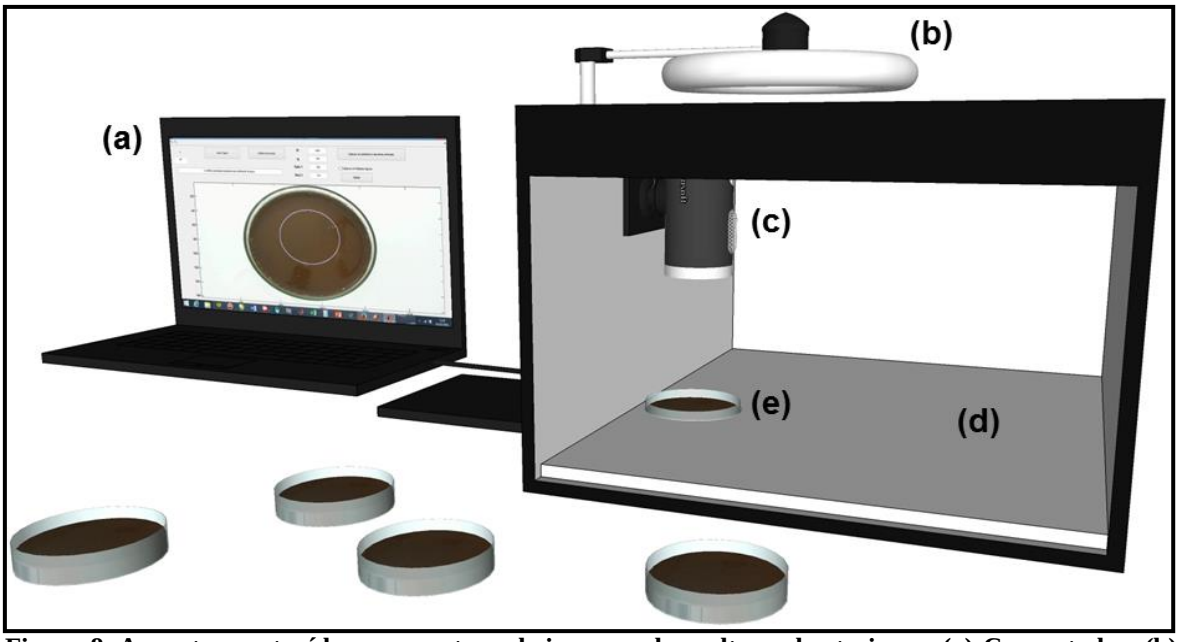

Figura 8. Aparato construído para captura de imagens das culturas bacterianas. (a) Computador, (b) lâmpada fluorescente, (c) webcam, (d) suporte de medida (e) compartimento da amostra.

3.3. Aquisição dos histogramas

Um total de 335 imagens (5 para cada amostra de bactéria) foram obtidas. A seguir, uma região que corresponde a 40% de cada imagem capturada foi selecionada e, em seguida, decomposto em histogramas de cor (RGB, HSI e intensidade de escala de cinza) usando o software Matlab 2010a. Histogramas médios para as 5 repetições de cada amostra de bactéria foram calculados.

3.4. Análise de dados

Os histogramas obtidos foram utilizados para a construção dos modelos de classificação multivariada por SIMCA, PCA-LDA, PLS-DA e SPA-LDA. As amostras foram divididas em conjuntos de treinamento (75%) e teste (25%) definidos através da aplicação do algoritmo de amostragem uniforme KS [32]. Diferentes técnicas de reconhecimento de padrões não supervisionada (PCA) e supervisionadas (SIMCA, PLS-

DA, PCA-LDA e SPA-LDA) foram avaliadas. A etapa de validação para todos os algoritmos foi realizada por meio de validação cruzada completa e as amostras de teste foram utilizadas apenas para a avaliação final dos dados e comparação de modelos de classificação [25].

Os algoritmos KS e SPA-LDA foram rodados no software Matlab® 2009b (Mathworks Inc.). As outras abordagens quimiométricas foram realizadas com a ferramenta para classificação *Classification toolbox* para Matlab® (versão 2.0), publicada pelo grupo de pesquisa Milano Chemometrics e QSAR [37], que pode ser encontrada no seguinte site: http://michem.disat.unimib.it/chm/.

4. RESULTADOS E DISCUSSÃO

Uma vez que o estado de crescimento das bactérias e a natureza das soluções nutriente foram mantidos sob as mesmas condições, os resultados dependem apenas da distribuição estatística dos pixels (histogramas de cor) como uma função da componente de cor em uma imagem digital.

Na **Figura 9a** são mostrados os histogramas médios das cinco classes de bactérias. Cada componente de cor dos modelos é composto de 256 tons, que são utilizados como informação analítica. A fim de verificar a influência relativa de cada cor foram selecionados seis diferentes modelos de cores empregando (a) escala de cinza, (b) RGB, (c) HSI, (d) escala de Cinza + RGB, (e) escala de cinza + HSI e (f) escala de cinza + RGB + HSI. Assim, a natureza multivariada complexa dos dados requer uma avaliação estatística usando técnicas de reconhecimento de padrões. Como os histogramas possuem variáveis com o valor zero, elas são removidas antes do tratamento quimiométrico, como pode ser visto na **Figura 9b**.

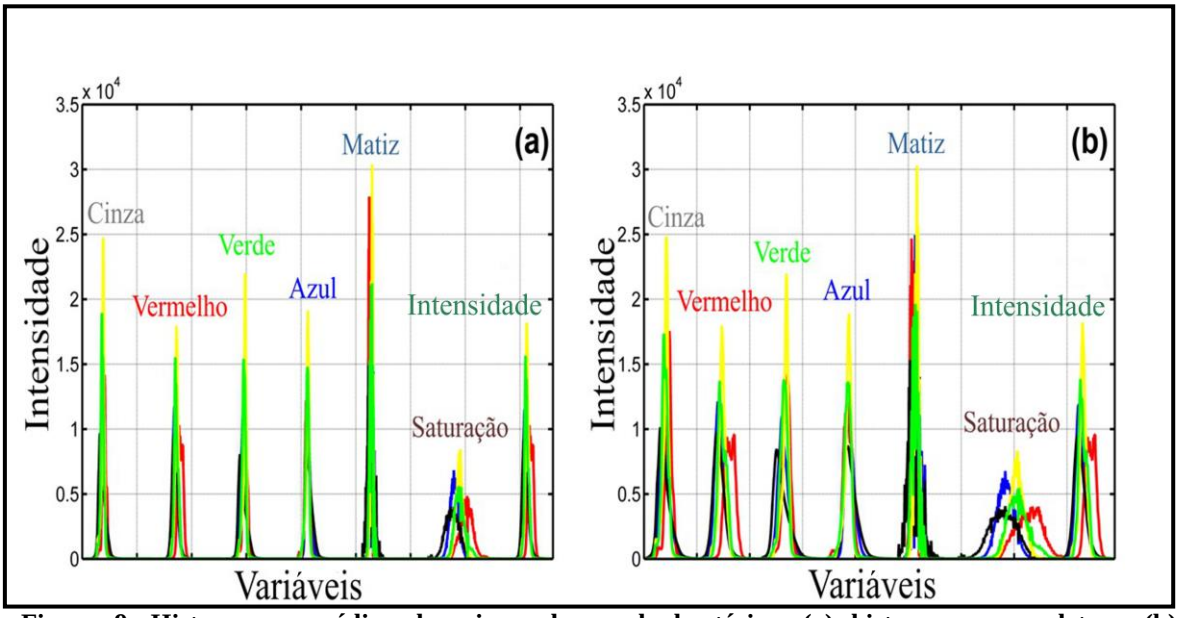

Figura 9. Histogramas médios das cinco classes de bactérias: (a) histograma completo e (b) histogramas após o tratamento quimiométrico.

4.1. Classificação

A construção dos modelos de classificação multivariada foi realizada utilizando um conjunto de treinamento (75% das amostras estudadas) selecionado pelo algoritmo KS e, em seguida, cada modelo foi validado usando a técnica de validação cruzada completa. Um conjunto de teste (25% das amostras estudadas) foi então usado apenas para a avaliação final dos dados e comparação entre os modelos de classificação. No **Apêndice 1** é apresentada a matriz de confusão contendo as atribuições das amostras do conjunto de teste para as cinco classes de bactérias estudadas usando SIMCA, PCA-LDA, PLS-DA e SPA-LDA. O desempenho dos modelos foi avaliado por meio de exatidão, que é definida como a proporção de amostras designadas corretamente no conjunto de teste dentro de suas classes respectivas. A **Tabela 1** apresenta o resumo da exatidão da classificação das amostras dos conjuntos de treinamento e teste para as cinco classes de bactérias estudadas usando SIMCA, PCA-LDA, PLS-DA e SPA-LDA.

Tabela 1. Resumo da exatidão da classificação das amostras dos conjuntos de treinamento e teste para as cinco classes de bactérias estudadas usando SIMCA, PCA-LDA, PLS-DA e SPA-LDA.

	Exatidão de Classificação (%)							
	SIMCA		PCA-LDA		PLS-DA		SPA-LDA	
Modelo de Cor	**VC**	**Teste**	**VC**	**Teste**	**VC**	**Teste**	**VC**	**Teste**
Cinza	44	76	64	76	66	70	92	65
RGB	56	82	74	88	74	94	94	100
HSI	70	94	76	88	78	82	96	82
Cinza+RGB	62	88	72	88	72	94	86	88
Cinza+HSI	64	88	78	94	78	94	100	88
Cinza+RGB+HSI	62	82	76	94	76	94	84	88

4.2. SIMCA

O melhor resultado obtido por SIMCA utilizou o modelo de cor HSI, que alcançou 70 e 94% de exatidão da classificação para os conjuntos de treinamento e teste, respectivamente. Para alcançar este resultado, o número ideal de componentes principais utilizados para cada classe foi: 1 PC por *Escherichia coli*, 3 PC para *Enterococcus faecalis*, 1 PC para *Streptococcus salivarius*, 2 PCs para *Streptococcus oralis* e 1 PC para *Staphylococcus aureus*.

4.3. PCA-LDA

O melhor resultado obtido por PCA-LDA foi atingido usando a combinação escala de cinza e HSI, chegando a 78 e 94% de exatidão de classificação para os conjuntos de treinamento e teste, respectivamente.

4.4. PLS-DA

Como em PCA-LDA, os melhores resultados PLS-DA foram obtidos utilizando a combinação escala de cinza + HSI, alcançando 78 e 94% de precisão de classificação nos conjuntos de treinamento e teste, respectivamente.

4.5. SPA-LDA

O melhor resultado SPA-LDA foi obtido utilizando RGB, que chegou a 94 e 100% de exatidão da classificação para os conjuntos de treinamento e teste, respectivamente. Na **Figura 10** é apresentado o histograma médio das amostras de bactérias estudadas, com as variáveis selecionadas por SPA.

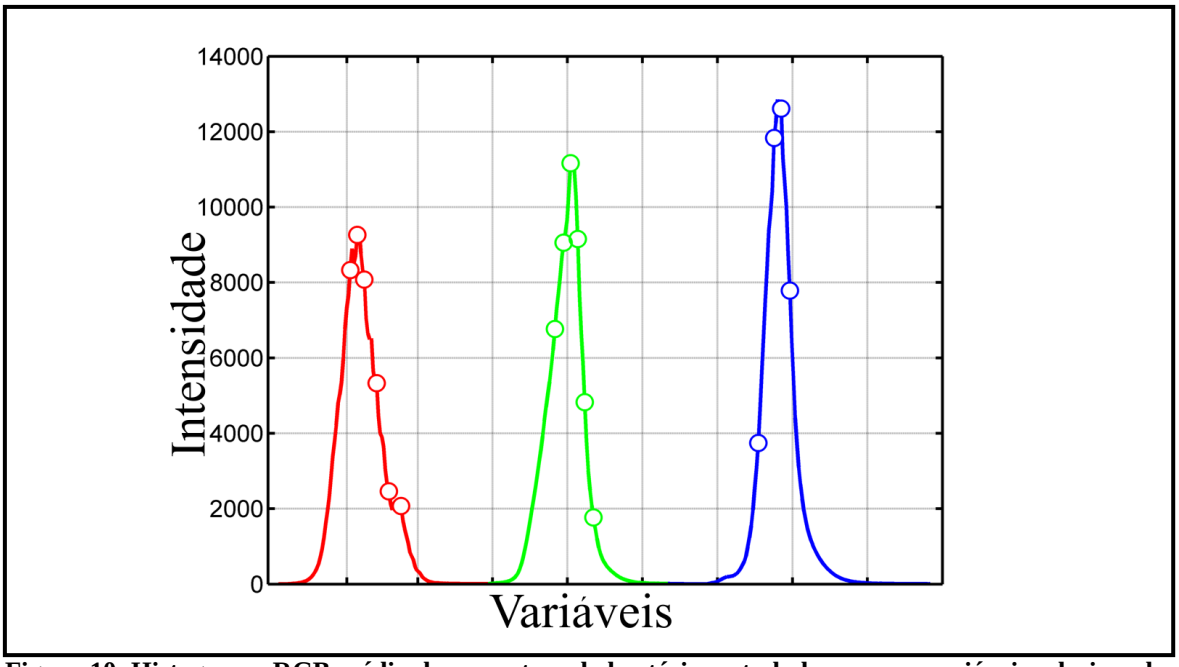

Figura 10. Histograma RGB médio das amostras de bactérias estudadas com as variáveis selecionadas por SPA.

Para ilustração, a **Figura 11** mostra a discriminação das amostras de bactérias usando o modelo RGB e as 16 variáveis selecionadas por SPA nas duas primeiras funções discriminantes (DFs) de Fisher.

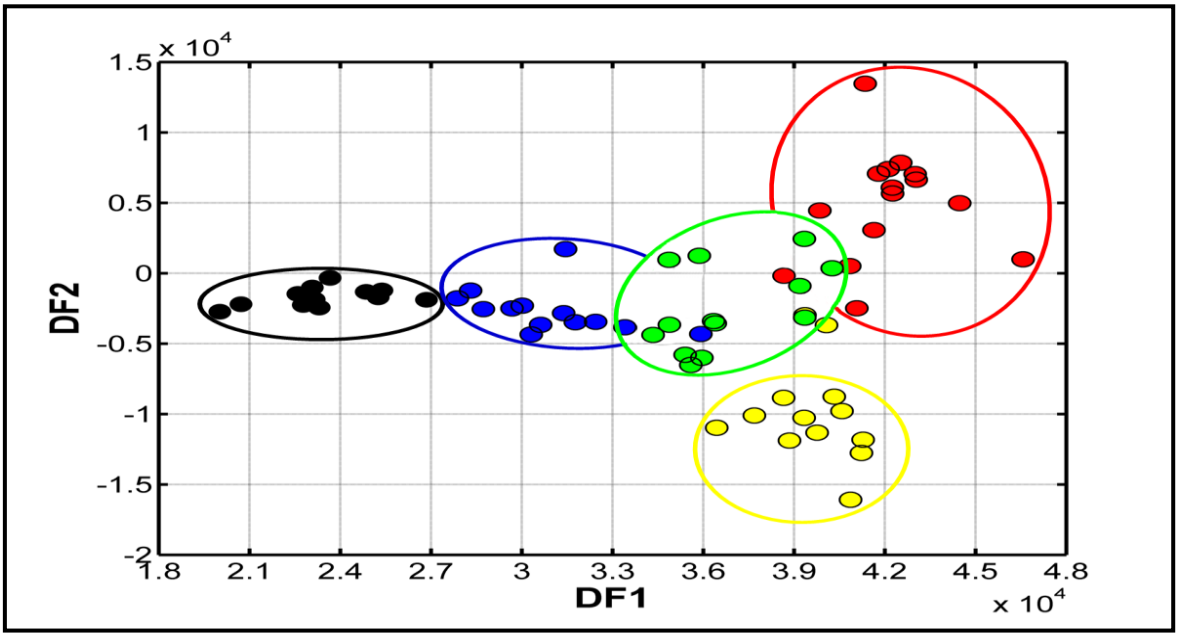

Figura 11. Funções discriminantes dos cinco tipos de bactérias utilizando o modelo de cor RGB e as variáveis selecionadas por SPA.

5. CONCLUSÃO

Este trabalho demonstra o uso de informações analíticas extraídas de histogramas de cor gerados a partir de imagens digitais e técnicas de reconhecimento de padrões supervisionadas para a classificação de cinco diferentes tipos de bactéria. Para este efeito, foram selecionados diferentes modelos de cor e suas combinações (escala de cinza, RGB, HSI, escala de cinza + RGB, escala cinza e escala cinza + RGB + HSI) e classificadores multivariados (SIMCA, PCA-LDA, PLS-DA e SPA-LDA).

O melhor resultado de classificação foi obtido usando RGB e SPA-LDA, que chegou a 94 e 100% de exatidão da classificação para os conjuntos de treinamento e teste, respectivamente. Este resultado é extremamente positivo do ponto de vista de análises microbiológicas e clínicas de rotina, porque evita a identificação de bactérias com base na identificação fenotípica do organismo causador usando coloração de Gram, a cultura das cepas e os testes bioquímicos. Portanto, o método proposto apresenta vantagens inerentes, promovendo uma alternativa mais simples, rápida e de baixo custo para a identificação de bactérias. Contudo, para garantir a generalização da metodologia proposta, deve ser incrementado um maior número de amostras de bactérias em estudos futuros.

Referências

[1] van der Merwe RG, van Helden PD, Warren RM, Sampson SL, van Pittius NCG (2014), Phage-based detection of bacterial pathogens. Analyst, Doi:10.1039/C4AN00208C

[2] Hossain Z (2014) In: Motarjemi Y (ed), Encyclopedia of Food Safety, 1st edn. Academic Press, San Diego.

[3] Codina MG, de Cueto M, Vicente D, Echevarría JE, Prats G (2011) Microbiological diagnosis of central nervous system infections. Enferm Infecc Microbiol Clin 29:127–134.

[4] Guibet F, Amiel C, Cadot P, Cordevant C, Desmonts MH, Lange M, Marecat A, Travert J, Denis C, Mariey L (2003) Discrimination and classification of *Enterococci* by Fourier transform infrared (FT-IR) spectroscopy. Vib Spectrosc 33:133–142.

[5] Schleifer KH (2009) Classification of Bacteria and *Archaea*: Past, present and future. Syst Appl Microbiol 32:533–542.

[6] Xiao D, Zhao F, Lv M, Zhang H, Zhang Y, Huang H, Su P, Zhang Z, Zhang J (2012) Rapid identification of microorganisms isolated from throat swab specimens of community-acquired pneumonia patients by two MALDI-TOF MS systems. Diagn Micr Infec Dis 73:301–307.

[7] Nakai S, Wang ZH, Dou J, Nakamura S, Ogawa M, Nakai E, Vanderstoep J (1999) Gas chromatography/principal component similarity system for detection of *E. coli* and *S. aureus* contaminating salmon and hamburger. J Agric Food Chem 47:576–583.

[8] Li D, Truong TV, Bills TM, Holt BC, Van Derwerken DN, Williams JR, Acharya A, Robison RA, Tolley HD, Lee ML (2012) GC/MS method for positive detection of *Bacillus anthracis* endospores. Anal Chem 84:1637–1644.

[9] Hantula J, Kurki A, Vuoriranta P, Bamford DH (1991) Rapid classification of bacterial strains by SDS-polyacrylamide gel electrophoresis: population dynamics of the dominant dispersed phase bacteria of activated sludge. Appl Microbiol Biotechnol 34:551–555.

[10] Veloo ACM, Erhard M, Welker M, Welling GW, Degener JE (2011) Identification of Gram-positive anaerobic cocci by MALDI-TOF mass spectrometry. Syst Appl Microbiol 34:58–62.

[11] Beier BD, Quivey Jr RG, Berger AJ (2010) Identification of different bacterial species in biofilms using confocal Raman microscopy. J Biomed Opt 15:066001 doi:10.1117/1.3505010.

[12] Oust A, Møretrø T, Kirschner C, Narvhus JA, Kohler A (2004) FT-IR spectroscopy for identification of closely related lactobacilli. J Microbiol Methods 59:149–162.

[13] Preisner O, Lopes JA, Menezes JC (2008) Uncertainty assessment in FT-IR spectroscopy based bacteria classification models. Chemom Intell Lab Syst 94:33–42.

[14] Marques AS, de Melo MCN, Cidral TA, de Lima KMG (2014) Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: A case study. J Microbiol Methods 98:26–30.

[15] Giana HE, Silveira Jr L, Zângaro RA, Pacheco MTT (2003) Rapid identification of bacterial species by fluorescence spectroscopy and classification through principal components analysis. J Fluoresc 13:489–493.

[16] Sohn M, Himmelsbach DS, Barton FE, Fedorka-Cray PJ (2009) Fluorescence spectroscopy for rapid detection and classification of bacterial pathogens. Appl Spectrosc 63:1251–1255.

[17] Kim SW, Ban SH, Ahn CY, Oh HM, Chung H, Cho SH, Park YM, Liu JR(2006) Taxonomic discrimination of cyanobacteria by metabolic fingerprinting using proton nuclear magnetic resonance spectra and multivariate statistical analysis. J Plant Biol 49:271–275.

[18] Dubuisson M-P, Jain AK, Jain MK (1994) Segmentation and classification of bacterial culture images. J Microbiol Methods 19:279–295.

[19] Kumar S, Mittal GS (2008) Geometric and optical characteristics of five microorganisms for rapid detection using image processing. Biosyst Eng 99:1–8.

[20] Huff K, Aroonnual A, Littlejohn AE, Rajwa B, Bae E, Banada PP, Patsekin V, Hirleman ED, Robinson JP, Richards GP, Bhunia AK (2012) Light-scattering sensor for real-time identification of *Vibrio parahaemolyticus*, *Vibrio vulnificus* and *Vibrio cholerae* colonies on solid agar plate. Microb Biotechnol 5:607–620.

[21] Suchwałko A, Buzalewicz I, Wieliczko A, Podbielska H (2013) Bacteria species identification by the statistical analysis of bacterial colonies Fresnel patterns. Opt Express 21:11322–11337.

[22] Diniz PHGD, Dantas HV, Melo KDT, Barbosa MF, Harding DP, Nascimento ECL, Pistonesi MF, Band BSF, Araújo MCU (2012) Using a simple digital camera and SPA-LDA modeling to screen teas. Anal Methods 4:2648–2652.

[23] Domínguez MA, Diniz PHGD, Di Nezio MS, Araújo MCU, Centurión ME (2014) Geographical origin classification of Argentinean honeys using a digital image-based flow-batch system. Microchem J 112:104–108.

[24] Milanez KDTM, Pontes MJC (2014) Classification of edible vegetable oil using digital image and pattern recognition techniques. Microchem J 113:10–16.

[25] Soares SFC, Gomes AA, Galvão Filho AR, Araújo MCU, Galvão RKH (2013) The successive projections algorithm. Trends Anal Chem 42:84–98.

[26] Tortora, G J, Funke, B R, Case, C L (2012), Microbiologia, 10th edn, Artimed, Brasil.

[27] Daniel, G. B. (2009) Digital imaging. Veterinary Clinics of North America: Small Animal Practice, 39:667–676.

[28] Gonzalez, R C, Woods, R E (2008). Digital Image Processing, 3rd ed, Pearson, Estados Unidos da América.

[29] Foley, J. D., Van Dam, A., Feiner, S. K., Hughes, J. F. (1990) Computer graphics, principles and practice. Reading: Addison-Wesley.

[30] Insausti, M, Romano C, Pistonesi, M F, Band, B S F, (2013) Simultaneous determination of quality parameters in biodiesel/diesel blends using synchronous fluorescence and multivariate analysis. Microchem J 10:832-37.

[31] Kenari, SLD, (2011) Alemzadeh, Maghsodi, Production of l-asparaginase from *Escherichia coli* ATCC 11303: Optimization by response surface methodology. Food Bioprod Process 89:315–321.

[32] Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148.

[33] Diniz, P H G D. Novas estratégias para classificação simultânea do tipo e origem geográfica de chás. 2013. 148 f. Tese de Doutorado. Universidade Federal da Paraíba, João Pessoa-PB, Brasil.

[34] Brereton, R.G. (2003) Chemometrics: data analysis for the laboratory and chemical plant. Ed. Wiley. University of Bristol, UK.

[35] Branden,K.V., (2005) Hubert,M. Robust classification in high dimensions based on the SIMCA Method. Chemom Intell Lab Syst 79:10-21.

[36] Tominaga, Y. (1999) Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. Chemom Intell Lab Syst 49:105–115.

[37] Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods 5:3790–3798.

[38] Agência Nacional de Vigilância Sanitária (2010) Farmacopeia Brasileira, Brasil.

[39] Clinical and Laboratory Standards Institute (2005) CLSI document M100-S15, EUA.

Apêndice 1 Atribuição das amostras de conjunto de teste para as cinco classes de bactérias estudadas usando SIMCA, PCA-LDA, PLS-DA, e SPA-LDA

	SIMCA					PCA-LDA					PLS-DA					SPA-LDA				
	EC	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**
Cinza																				
EC	4	0	0	0	0	4	0	0	2	0	4	0	0	0	0	4	0	0	1	0
EF	0	2	0	0	0	0	2	1	0	0	0	2	0	3	0	1	1	1	2	1
SS	0	0	2	0	0	0	0	2	0	1	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	2	0	0	1	0	2	0	0	1	0	1	1
SA	0	0	0	0	3	0	0	0	0	3	0	0	0	1	2	0	0	0	0	3
RGB																				
EC	2	1	0	1	0	4	0	0	2	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	3	0	0	2	0	4	0	0	1	0	5	0	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	2	0	1	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3
HSI																				
EC	3	0	0	0	1	4	0	0	2	0	4	0	0	0	0	4	0	0	1	0
EF	0	5	0	0	0	0	3	0	0	2	0	3	0	0	2	0	3	0	0	2
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	0	3	0	0	0	0	3	1	0	0	0	2	1	0	0	0	2

Cinza+RGB																				
EC	3	0	0	0	1	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	1	0	4	0	0	1	0	4	1	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	1	0	2	0	0	0	0	3	0	0	1	0	2	0
SA	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	2	0	0	3
Cinza+HSI																				
EC	3	0	0	0	1	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	0	0	5	0	0	0	0	4	1	0	1
SS	0	0	2	0	0	0	0	2	0	1	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	0	3	0	0	0	0	3	1	0	0	0	2	1	0	0	0	2
Cinza+RGB+HSI																				
EC	3	0	0	2	0	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	1	0	4	0	1	0	0	5	0	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	1	2	0	0	0	0	3	0	0	0	0	3	1	2	1	0	1

EC: ***Escherichia coli*****; EF:** ***Enterococcus faecalis;*** **ES:** ***Streptococcus salivarius;*** **SO,** ***Streptococcus oralis*****; SA:** ***Staphylococcus aureus.***

Anexo 1: Artigo referente ao trabalho de conclusão de curso publicado na revista internacional ABC Springer Qualis capes 2014 A2.

Anal Bioanal Chem (2014) 406:5989–5995
DOI 10.1007/s00216-014-8015-1

RESEARCH PAPER

Using color histograms and SPA-LDA to classify bacteria

Valber Elias de Almeida · Gean Bezerra da Costa · David Douglas de Sousa Fernandes · Paulo Henrique Gonçalves Dias Diniz · Deysiane Brandão · Ana Claudia Dantas de Medeiros · Germano Véras

Received: 13 May 2014 / Revised: 28 June 2014 / Accepted: 30 June 2014 / Published online: 15 July 2014

Abstract In this work, a new approach is proposed to verify the differentiating characteristics of five bacteria (*Escherichia coli*, *Enterococcus faecalis*, *Streptococcus salivarius*, *Streptococcus oralis*, and *Staphylococcus aureus*) by using digital images obtained with a simple webcam and variable selection by the Successive Projections Algorithm associated with Linear Discriminant Analysis (SPA-LDA). In this sense, color histograms in the red–green–blue (RGB), hue-saturation-value (HSV), and grayscale channels and their combinations were used as input data, and statistically evaluated by using different multivariate classifiers (Soft Independent Modeling by Class Analogy (SIMCA), Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA), Partial Least Squares Discriminant Analysis (PLS-DA) and Successive Projections Algorithm-Linear Discriminant Analysis (SPA-LDA)). The bacteria strains were cultivated in a nutritive blood agar base layer for 24 h by following the Brazilian Pharmacopoeia, maintaining the status of cell growth and the nature of nutrient solutions under the same conditions. The best result in classification was obtained by using RGB and SPA-LDA, which reached 94 and 100 % of classification accuracy in the training and test sets, respectively. This result is extremely positive from the viewpoint of routine clinical analyses, because it avoids bacterial identification based on phenotypic identification of the causative organism using Gram staining, culture, and biochemical proofs. Therefore, the proposed method presents inherent advantages, promoting a simpler, faster, and low-cost alternative for bacterial identification.

Keywords Bacteria · Digital images · Classification · Linear discriminant analysis · Successive projections algorithm

V. E. de Almeida · G. B. da Costa · D. D. de Sousa Fernandes · P. H. Gonçalves Dias Diniz · G. Véras (✉)
Centro de Ciência e Tecnologia, Laboratório de Química Analítica e Quimiometria (LQAQ), Universidade Estadual da Paraíba, CEP 58.429-500 Campina Grande, Paraíba, Brazil
e-mail: germano.veras@pq.cnpq.br

D. Brandão · A. C. D. de Medeiros
Centro de Ciências Biológicas e da Saúde, Laboratório de Desenvolvimento e Ensaios de Medicamentos (LABDEM), Universidade Estadual da Paraíba, CEP 58.429-500 Campina Grande, Paraíba, Brazil

Introduction

Infectious diseases caused by pathogenic bacteria are one of the factor most responsible for human and animal mortality around the world [1, 2]. The rapid diagnosis of these diseases with their respective precursors is a primordial step for the immediate initiation of proper treatment, increasing their effectiveness especially in patients with weakened immune systems [3]. The traditional methodologies used in the laboratories of clinical microbiology for bacteria identification are based on the phenotypic characteristics, biochemical, and metabolic properties of the microorganism [4]. In this sense, the formation of patterns in the growth of bacterial colonies in culture media, the morphology of the colonies, the Gram coloring, and some biochemical proofs are evaluated. However, these methodologies require comprehensive knowledge about the individual characteristics of each species of microorganism, besides being very laborious and slow, requiring more than 48 h for expert appraisal. Moreover, some strains exhibit unique biochemical characteristics that do not fit into patterns that have been used as a characteristic of any known genus and species [5, 6].

In order to circumvent the inconveniences described before, some methodologies for identification and classification of bacteria have been reported in the literature. They include gas chromatography [7, 8], capillary electrophoresis [9], mass spectroscopy [6, 10], Raman spectroscopy [11], infrared

spectroscopy [12–14], fluorescence [15, 16], and nuclear magnetic resonance [17]. Despite the ability of these instrumental analytical techniques to obtain reliable and accurate results in the identification and classification of microorganisms, the use of digital images presents some advantages, since it has low-cost equipment, avoids the manipulation of the Petri plate where microorganisms are grown, and does not require the knowledge of a skillful microbiologist. In this sense, some works concerning image analysis have been proposed in the literature [18–21].

Dubuisson et al. [18] developed an approach using segmentation of digital images for classification of *Methanospirillum hungatei* and *Methanosarcina mazei* based on their shapes. The *M. hungatei* culture were always correctly identified, meanwhile the results for *M. mazei* culture were not as accurate.

Kumar and Mittal [19] obtained geometrical and optical parameters using fluorescence microscopy and image analysis for identification of five microorganisms. The images of *Bacillus thuringiensis*, *Escherichia coli K12*, *Lactobacillus brevis*, *Listeria innocua*, and *Staphylococcus epidermidis* stained with two fluorescent dyes were captured using a CCD (charge-coupled device) camera attached to a light microscope. Fluorescence emission (gray-level intensity) from *B. thuringiensis* was the highest compared to other microbes, and the emission from *L. brevis* was the lowest. Mean 10 percentile values of image histograms of *L. innocua* and *S. epidermidis* were significantly different from that of *L. brevis*. Using 99 percentile values, *B. thuringiensis* can be differentiated from the remaining microbes, and *E. coli* can also be differentiated from *L. brevis* and *S. epidermidis*.

Huff et al. [20] used a light-scattering sensor for real-time identification of *Vibrio parahaemolyticus*, *Vibrio vulnificus*, and *Vibrio cholerae* colonies on solid agar plate. The colonies were illuminated by a 635 nm laser beam and scatter-image signatures were acquired using a CCD camera. A pattern recognition system provided a classification accuracy of 99 %. The proposed methodology successfully detected *V. cholerae*, *V. parahaemolyticus*, and *V. vulnificus* in oyster or water samples in 18 h even in the presence of other vibrios or other bacteria.

Suchwalko et al. [21] identified bacteria species (*Salmonella enteritidis*, *Staphylococcus aureus*, *Staphylococcus intermedius*, *E. coli*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, and *Citrobacter freundii*) based on their colonies' Fresnel diffraction patterns recorded in an optical system with converging spherical wave illumination. The proposed method used image processing and statistical analysis based on feature extraction, feature selection, and classification methods, reaching 98 % identification accuracy in 36 h from sample acquiring.

All approaches mentioned before uses geometrical parameters and/or feature extraction from the images of the bacteria. On the other hand, color histograms describe the statistical distribution of the pixels in a digital image as a function of the recorded color component, and not a feature or a physical-chemical behavior directly [22]. Color histograms have been successfully used as input data for classification of teas [22], honeys [23], and edible vegetable oils [24].

In this work is proposed a new approach to verify the differentiating characteristics of five bacteria (*E. coli*, *Enterococcus faecalis*, *Streptococcus salivarius*, *Streptococcus oralis*, and *S. aureus*) by using digital images obtained with a simple webcam and variable selection by the Successive Projections Algorithm associated with Linear Discriminant Analysis (SPA-LDA) [25]. In this sense, color histograms in the red–green–blue (RGB), hue-saturation-value (HSV), and grayscale channels were used as input data, and statistically evaluated by using supervised pattern recognition techniques such as Soft Independent Modeling by Class Analogy (SIMCA), Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA), Partial Least Squares Discriminant Analysis (PLS-DA), and Successive Projections Algorithm associated with Linear Discriminant Analysis (SPA-LDA).

Materials and methods

Samples

American Type Culture Collection (ATCC) standard strains of five differing types of freeze-dried bacteria were released by the Oswaldo Cruz Foundation, Brazil. For this study, *E. coli* (14 samples), *E. faecalis* (15 samples), *S. salivarius* (13 samples), *S. oralis* (13 samples), and *S. aureus* (12 samples) were selected.

The bacterial inoculum was standardized by following the Brazilian Pharmacopoeia [26], which is based on the methodology of the Clinical and Laboratory Standards Institute [27]. Firstly, the bacterial samples were prepared with brain heart infusion (BHI) broth until 85 % transmittance at 630 nm in a UV–Vis spectrophotometer Biospectro, model SP22 is achieved, in order to obtain a bacterial preparation with a final concentration of 106 CFU (colony forming unit) mL^{-1}. Then, 1 mL of this suspension was diluted with 9 mL in a saline solution of NaCl 0.9 % m v^{-1}. A base layer was then produced by addition of 20 mL of a nutritive blood agar over the Petri plate. After the hardening of the blood agar, 5 mL of the standardized inoculum was added over the base layer, waiting its re-hardening. The plates were then incubated at 37 °C for 24 h. This methodology is summarized in Fig. 1.

Instrumentation

Figure 2 shows the apparatus built for image acquisition of the bacteria cultures. A box with dimensions sized 30 cm×

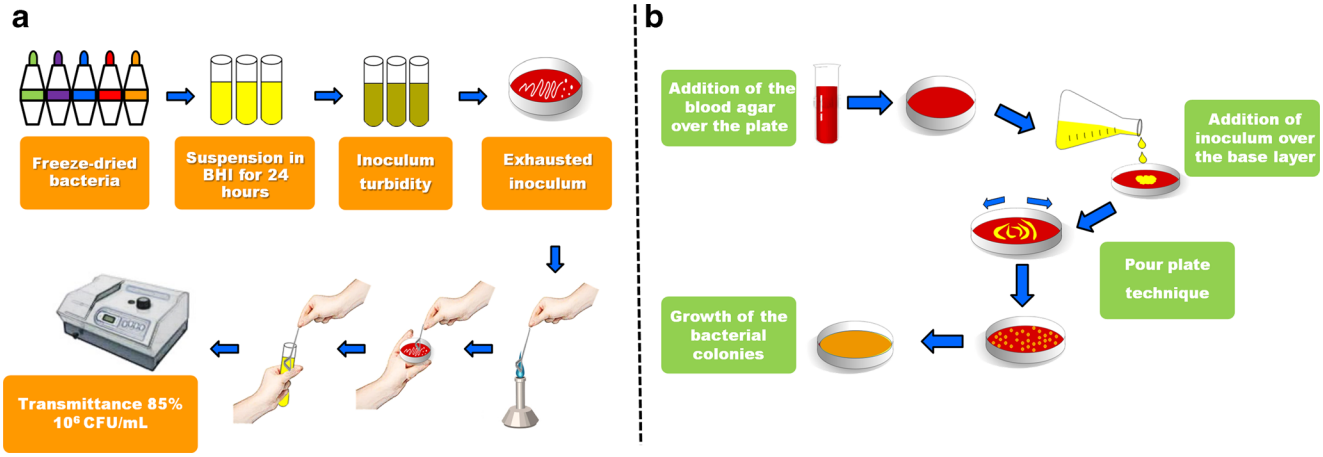

Fig. 1 **a** Standardization of the bacterial inoculum and **b** preparation of bacteria cultures on a nutritive agar plate

22 cm×23 cm was built and internally covered with white office paper in order to avoid external light interferences and light reflection into the box, ensuring the uniformity to the captured images [24]. A Webcam, Microsoft® model Lifecam Cinema, with 7.1 megapixels, was set above the sample holder vertically. The distance between the camera and the sample holder was 15 cm, and between the illumination and the sample holder was 25 cm. The webcam was placed in the center of a circular fluorescent lamp of 8 W to homogenize the internal illumination, as shown in Fig. 2.

Acquisition of the histograms

A total of 335 images (5 for each bacterium sample) were obtained. Then, a region corresponding to 40 % of each captured image was selected and then decomposed into color histograms (RGB, HSV, and grayscale intensity) by using Matlab 2010a software. Mean histograms from the five replicates for each bacterium sample were calculated and were then used as analytical information.

Data analysis

The obtained histograms were used for building the multivariate classification models by SIMCA, PCA-LDA, PLS-DA, and SPA-LDA. The samples were divided into training (75 %), and test (25 %) sets by applying the Kennard-Stone (KS) uniform sampling algorithm [28]. Differing unsupervised (PCA) and supervised (SIMCA, PLS-DA, PCA-LDA, and SPA-LDA) pattern recognition were evaluated. The validation step for all algorithms was performed using full cross-validation. The test samples were only used for the final data evaluation and comparison of the classification models [25]. The performance of the models was evaluated by means of the accuracy, which is defined as the ratio of correctly assigned samples in the test set into their respective classes.

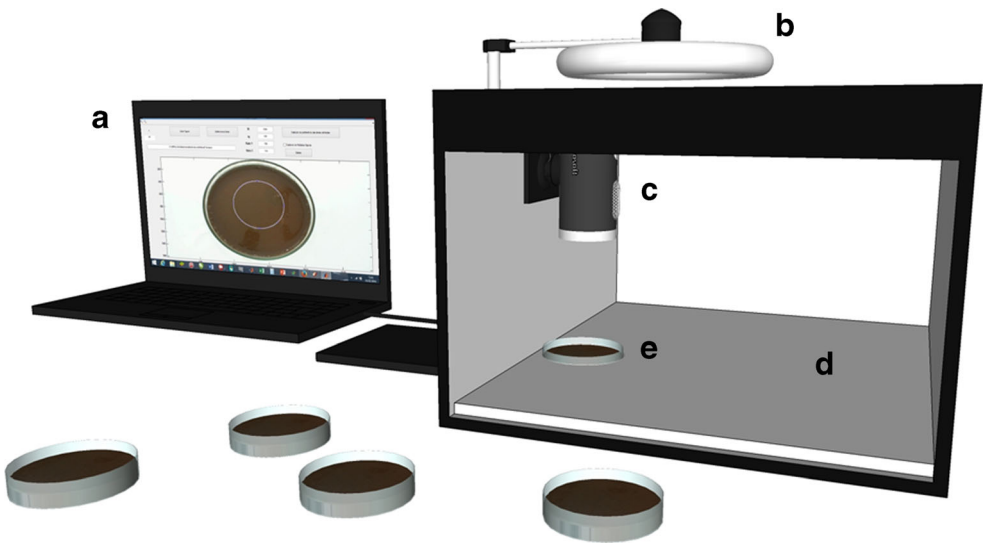

Fig. 2 Apparatus built for image capturing of the bacteria cultures. (*a*) Notebook, (*b*) fluorescent lamp, (*c*) webcam, (*d*) sample holder, (*e*) box

The KS and SPA-LDA algorithms were performed with Matlab® 2009b (Mathworks Inc.) software. The other chemometric approaches were performed by using the Classification toolbox for Matlab® (version 2.0) released by Milano Chemometrics and QSAR Research Group [29]. It is found in the following webpage: http://michem.disat.unimib.it/chm/.

Results and discussion

Since the status of cell growth and the nature of nutrient solutions were under the same conditions, the results depend only on the statistical distribution of the pixels (color histograms) as a function of the recorded color component in a digital image. Figure 3a shows the mean histograms of the five classes of bacteria. Each color component of the grayscale, RGB, and HSV systems is composed of 256 tones, which are used as analytical information. In order to check for each color's relative influence, different color systems and their combinations were selected: (a) grayscale, (b) RGB, (c) HSV, (d) grayscale+RGB, (e) grayscale+HSV, and (f) grayscale+RGB+HSV. Therefore, the complex and multivariate nature of the data requires a statistical evaluation by using supervised pattern recognition techniques. Since the histograms have variables with zero value, they are removed before the chemometric treatment, as can be seen in Fig. 3b.

Classification

Table 1 shows the assignation in the test set samples into the five studied classes of bacteria using SIMCA, PCA-LDA, PLS-DA, and SPA-LDA. Table 2 presents the summary of the classification accuracy in the test set samples into the five studied classes of bacteria using SIMCA, PCA-LDA, PLS-DA, and SPA-LDA.

SIMCA

SIMCA is a class modeling technique, in which the final classification model consists of a collection of PCA models, one for each class. Then, a new object is assigned by comparing the distances of the object from the class models. The best SIMCA result was obtained by using HSV, reaching 70 and 94 % of classification accuracy in the training and test sets, respectively. To achieve this result, the optimal number of principal components used for each class was: 1 PC for *E. coli*, 3 PC for *E. faecalis*, 1 PC for *S. salivarius*, 2 PCs for *S. oralis*, and 1 PC for *S. aureus*.

PCA-LDA

The linear discriminant analysis is performed by maximizing the between-class separability while minimizing their within-class variability based on a distance calculation. Due to mathematical limitations, LDA models require a step of variable reduction. This can be done by using the scores of PCA as input data, because linear combinations of the original variables called principal components (PCs) are uncorrelated. The best PCA-LDA result was attained by using grayscale+HSV, reaching 78 and 94 % of classification accuracy in the training and test sets, respectively.

PLS-DA

PLS-DA is based on the PLS2 algorithm combined with the discriminant analysis, which searches for latent variables with a maximum covariance with the categorical variables (*Y*). The new object is then assigned to the class with the maximum value in the *Y* vector or, alternatively, a threshold between zero and one is determined for each class. Like in PCA-LDA, the best PLS-DA result were obtained by using grayscale+HSV, achieving 78 and

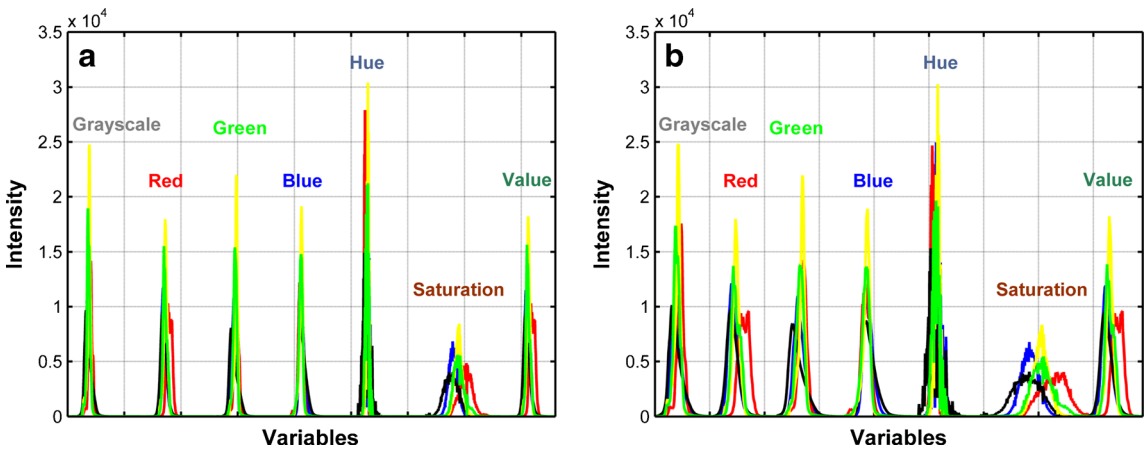

Fig. 3 Mean histograms of the five classes of bacteria: **a** full-histograms and **b** histograms used in the chemometric treatment

Table 1 Assignation of the test set samples into the five studied bacteria classes using SIMCA, PCA-LDA, PLS-DA, and SPA-LDA

	SIMCA					PCA-LDA					PLS-DA					SPA-LDA				
	EC	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**	**EC**	**EF**	**SS**	**SO**	**SA**
Grayscale																				
EC	4	0	0	0	0	4	0	0	2	0	4	0	0	0	0	4	0	0	1	0
EF	0	2	0	0	0	0	2	1	0	0	0	2	0	3	0	1	1	1	2	1
SS	0	0	2	0	0	0	0	2	0	1	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	2	0	0	1	0	2	0	0	1	0	1	1
SA	0	0	0	0	3	0	0	0	0	3	0	0	0	1	2	0	0	0	0	3
RGB																				
EC	2	1	0	1	0	4	0	0	2	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	3	0	0	2	0	4	0	0	1	0	5	0	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	2	0	1	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3
HSV																				
EC	3	0	0	0	1	4	0	0	2	0	4	0	0	0	0	4	0	0	1	0
EF	0	5	0	0	0	0	3	0	0	2	0	3	0	0	2	0	3	0	0	2
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	0	3	0	0	0	0	3	1	0	0	0	2	1	0	0	0	2
Grayscale+RGB																				
EC	3	0	0	0	1	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	1	0	4	0	0	1	0	4	1	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	1	0	2	0	0	0	0	3	0	0	1	0	2	0
SA	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	2	0	0	3
Grayscale+HSV																				
EC	3	0	0	0	1	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	0	0	5	0	0	0	0	4	1	0	1
SS	0	0	2	0	0	0	0	2	0	1	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	0	3	0	0	0	0	3	1	0	0	0	2	1	0	0	0	2
Grayscale+RGB+HSV																				
EC	3	0	0	2	0	4	0	0	0	0	4	0	0	0	0	4	0	0	0	0
EF	0	5	0	0	0	0	4	0	0	1	0	4	0	1	0	0	5	0	0	0
SS	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0
SO	0	1	0	2	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0
SA	0	0	0	1	2	0	0	0	0	3	0	0	0	0	3	1	2	1	0	1

EC E. coli; *EF E. faecalis*; *ES S. salivarius*; *SO S. oralis*; *SA S. aureus*

Table 2 Summary of the classification accuracy in the test set samples into the five studied classes of bacteria using SIMCA, PCA-LDA, PLS-DA, and SPA-LDA

	Classification accuracy (%)							
	SIMCA		PCA-LDA		PLS-DA		SPA-LDA	
Color models	CV	Test	CV	Test	CV	Test	CV	Test
Gray	44	76	64	76	66	70	92	65
RGB	56	82	74	88	74	94	94	100
HSV	70	94	76	88	78	82	96	82
Grayscale+RGB	62	88	72	88	72	94	86	88
Grayscale+HSV	64	88	78	94	78	94	100	88
Grayscale+RGB+HSV	62	82	76	94	76	94	84	88

94 % of classification accuracy in the training and test sets, respectively.

SPA-LDA

SPA is an iterative forward selection method that solves collinearity problems by selecting variables whose information content is minimally redundant. The chains of variables are then sequentially evaluated based on a cost function; in the case of LDA, the *G cost* function [25] is used. The best SPA-LDA result was attained by using RGB, which reached 94 and 100 % of classification accuracy in the training and test sets, respectively. Figure 4 shows the mean histogram of the studied bacterium samples with the selected variables by using SPA. For illustration, Fig. 5 shows the discrimination of the bacteria samples by using RGB in the first two Fisher's discriminant functions. It is worth to highlight that SPA-LDA also achieved 100 % of classification accuracy in the training set by using grayscale+HSV. Therefore, SPA-LDA demonstrated to be the most suitable approach for the classification of bacteria samples into the five differing bacterium classes: *E. coli*, *E. faecalis*, *S. salivarius*, *S. oralis*, and *S. aureus*.

Conclusion

This paper demonstrates that the use of analytical information extracted from digital images generated color histograms and supervised pattern recognition techniques for classification of five differing bacterium types. For this purpose, three color systems (grayscale, RGB, and HSV) and their combinations with differing multivariate classifiers (SIMCA, PCA-LDA, PLS-DA, and SPA-LDA) were selected. The best result in classification was obtained by using RGB and SPA-LDA, which reached 94 and 100 % of classification accuracy in

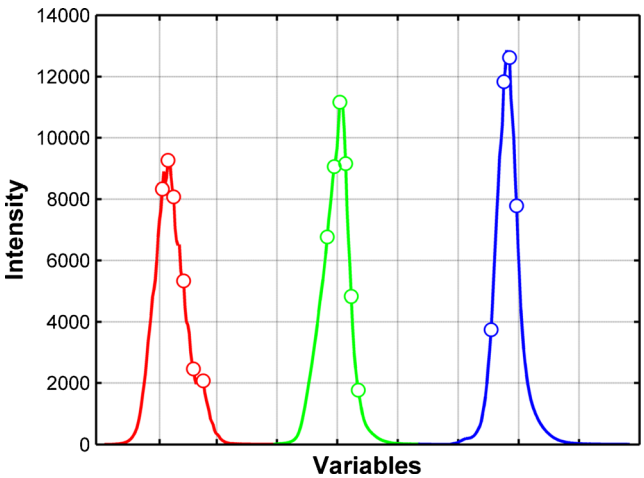

Fig. 4 Mean RGB histogram of the studied bacteria samples with the wavelengths selected by using SPA

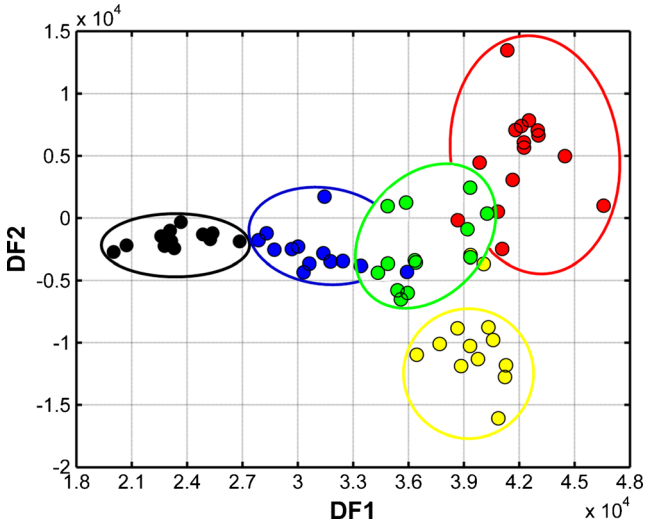

Fig. 5 Discrimination of the bacteria samples by using RGB in the first tow Fisher's discriminant functions

the training and test sets, respectively. This result is extremely positive from the viewpoint of routine clinical analyses, because it avoids bacterial identification based on phenotypic identification of the causative organism using Gram staining, culture, and biochemical methods. Therefore, the proposed method presents inherent advantages, promoting a simpler, faster, and low-cost alternative for bacterial identification. However, to guarantee any generalization of the proposed methodology, a larger (more varied) testing of bacterium samples using more types of bacteria must be implemented.

Acknowledgments The authors gratefully acknowledge the Universidade Estadual da Paraíba for the financial support. The authors thank also Capes and CNPq Brazil scholarships and research fellowships. Paulo Henrique Gonçalves Dias Diniz also thanks to Fundação de Apoio à Pesquisa do Estado da Paraíba – FAPESQ-PB.

References

1. van der Merwe RG, van Helden PD, Warren RM, Sampson SL, van Pittius NCG (2014) Phage-based detection of bacterial pathogens. Analyst. doi:10.1039/C4AN00208C
2. Hossain Z (2014) In: Motarjemi Y (ed) Encyclopedia of food safety, 1st edn. Academic, San Diego
3. Codina MG, de Cueto M, Vicente D, Echevarría JE, Prats G (2011) Microbiological diagnosis of central nervous system infections. Enferm Infecc Microbiol Clin 29:127–134
4. Guibet F, Amiel C, Cadot P, Cordevant C, Desmonts MH, Lange M, Marecat A, Travert J, Denis C, Mariey L (2003) Discrimination and classification of *Enterococci* by Fourier transform infrared (FT-IR) spectroscopy. Vib Spectrosc 33:133–142
5. Schleifer KH (2009) Classification of bacteria and *Archaea*: past, present and future. Syst Appl Microbiol 32:533–542
6. Xiao D, Zhao F, Lv M, Zhang H, Zhang Y, Huang H, Su P, Zhang Z, Zhang J (2012) Rapid identification of microorganisms isolated from throat swab specimens of community-acquired pneumonia patients by two MALDI-TOF MS systems. Diagn Microbiol Infect Dis 73:301–307
7. Nakai S, Wang ZH, Dou J, Nakamura S, Ogawa M, Nakai E, Vanderstoep J (1999) Gas chromatography/principal component similarity system for detection of *E. coli* and *S. aureus* contaminating salmon and hamburger. J Agric Food Chem 47:576–583
8. Li D, Truong TV, Bills TM, Holt BC, Van Derwerken DN, Williams JR, Acharya A, Robison RA, Tolley HD, Lee ML (2012) GC/MS method for positive detection of *Bacillus anthracis* endospores. Anal Chem 84:1637–1644
9. Hantula J, Kurki A, Vuoriranta P, Bamford DH (1991) Rapid classification of bacterial strains by SDS-polyacrylamide gel electrophoresis: population dynamics of the dominant dispersed phase bacteria of activated sludge. Appl Microbiol Biotechnol 34:551–555
10. Veloo ACM, Erhard M, Welker M, Welling GW, Degener JE (2011) Identification of Gram-positive anaerobic cocci by MALDI-TOF mass spectrometry. Syst Appl Microbiol 34:58–62
11. Beier BD, Quivey RG Jr, Berger AJ (2010) Identification of different bacterial species in biofilms using confocal Raman microscopy. J Biomed Opt 15:066001
12. Oust A, Møretrø T, Kirschner C, Narvhus JA, Kohler A (2004) FT-IR spectroscopy for identification of closely related lactobacilli. J Microbiol Methods 59:149–162
13. Preisner O, Lopes JA, Menezes JC (2008) Uncertainty assessment in FT-IR spectroscopy based bacteria classification models. Chemom Intell Lab Syst 94:33–42
14. Marques AS, de Melo MCN, Cidral TA, de Lima KMG (2014) Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: a case study. J Microbiol Methods 98:26–30
15. Giana HE, Silveira L Jr, Zângaro RA, Pacheco MTT (2003) Rapid identification of bacterial species by fluorescence spectroscopy and classification through principal components analysis. J Fluoresc 13:489–493
16. Sohn M, Himmelsbach DS, Barton FE, Fedorka-Cray PJ (2009) Fluorescence spectroscopy for rapid detection and classification of bacterial pathogens. Appl Spectrosc 63:1251–1255
17. Kim SW, Ban SH, Ahn CY, Oh HM, Chung H, Cho SH, Park YM, Liu JR (2006) Taxonomic discrimination of cyanobacteria by metabolic fingerprinting using proton nuclear magnetic resonance spectra and multivariate statistical analysis. J Plant Biol 49:271–275
18. Dubuisson M-P, Jain AK, Jain MK (1994) Segmentation and classification of bacterial culture images. J Microbiol Methods 19:279–295
19. Kumar S, Mittal GS (2008) Geometric and optical characteristics of five microorganisms for rapid detection using image processing. Biosyst Eng 99:1–8
20. Huff K, Aroonnual A, Littlejohn AE, Rajwa B, Bae E, Banada PP, Patsekin V, Hirleman ED, Robinson JP, Richards GP, Bhunia AK (2012) Light-scattering sensor for real-time identification of *Vibrio parahaemolyticus*, *Vibrio vulnificus* and *Vibrio cholerae* colonies on solid agar plate. Microb Biotechnol 5:607–620
21. Suchwałko A, Buzalewicz I, Wieliczko A, Podbielska H (2013) Bacteria species identification by the statistical analysis of bacterial colonies Fresnel patterns. Opt Express 21:11322–11337
22. Diniz PHGD, Dantas HV, Melo KDT, Barbosa MF, Harding DP, Nascimento ECL, Pistonesi MF, Band BSF, Araújo MCU (2012) Using a simple digital camera and SPA-LDA modeling to screen teas. Anal Methods 4:2648–2652
23. Domínguez MA, Diniz PHGD, Di Nezio MS, Araújo MCU, Centurión ME (2014) Geographical origin classification of Argentinean honeys using a digital image-based flow-batch system. Microchem J 112:104–108
24. Milanez KDTM, Pontes MJC (2014) Classification of edible vegetable oil using digital image and pattern recognition techniques. Microchem J 113:10–16
25. Soares SFC, Gomes AA, Galvão Filho AR, Araújo MCU, Galvão RKH (2013) The successive projections algorithm. Trends Anal Chem 42:84–98
26. Agência Nacional de Vigilância Sanitária (2010) Farmacopeia brasileira, Brasil
27. Clinical and Laboratory Standards Institute (2005) CLSI document M100-S15, EUA
28. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148
29. Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods 5:3790–3798