



COPPE/UFRJ

MINERAÇÃO DE DADOS CITOMÉTRICOS:
OBTENÇÃO DE CONHECIMENTO DE PADRÕES CELULARES PARA
OTIMIZAÇÃO DE PROCESSOS BIOTECNOLÓGICOS

Ana Reis de Figueiredo

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

Orientadores: Nelson Francisco Favilla Ebecken
Gilberto Carvalho Pereira

Rio de Janeiro
Março de 2010

MINERAÇÃO DE DADOS CITOMÉTRICOS:
OBTENÇÃO DE CONHECIMENTO DE PADRÕES CELULARES PARA
OTIMIZAÇÃO DE PROCESSOS BIOTECNOLÓGICOS

Ana Reis de Figueiredo

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

Examinada por:

Prof. Nelson Francisco Favilla Ebecken, D. Sc.

Dr. Gilberto Carvalho Pereira, D. Sc.

Prof. Luiz Bevilacqua, Ph.D.

Dr. Ernesto Hofer, L.D

Prof. Orlando Martins, Ph. D.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2010

Figueiredo, Ana Reis de

Mineração de Dados Citométricos: Obtenção de Conhecimento de Padrões Celulares para Otimização de Processos Biotecnológicos/ Ana Reis de Figueiredo – Rio de Janeiro: UFRJ/COPPE, 2010.

XII, 97 p.: il.; 29,7 cm.

Orientadores: Nelson Francisco Favilla Ebecken

Gilberto Carvalho Pereira

Tese (doutorado) – UFRJ/COPPE/ Programa de Engenharia Civil, 2010.

Referências Bibliográficas: p. 87-97.

1. Mineração de dados. 2. Citometria de fluxo. 3. Padrões Celulares. I. Figueiredo, Ana Reis de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original.

Albert Einstein

À memória do meu pai.

Agradecimentos

Primeiramente gostaria de agradecer ao meu orientador Prof. Nelson F. F. Ebecken pela oportunidade em desenvolver este trabalho, pela orientação e confiança em mim depositada.

Ao meu co-orientador Dr. Gilberto Carvalho Pereira agradeço pela orientação, pelo entusiasmo e dedicação durante todas as etapas do desenvolvimento desta tese.

Aos professores do Programa de Engenharia Civil da COPPE/UFRJ que durante a minha formação me ajudaram com seus conhecimentos.

Ao Dr. Ricardo Vieira do Depto. de Bioquímica Médica do CCS/UFRJ pelo apoio durante a execução da fase experimental da tese.

A todos os colegas do Programa de Engenharia Civil agradeço pela ajuda e boa convivência.

Aos funcionários do departamento pelo apoio dispensado.

A minha mãe pela ajuda durante todos os meus anos de vida e em especial estes últimos quatro anos de tese.

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro através da bolsa de estudos a mim concedida.

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

MINERAÇÃO DE DADOS CITOMÉTRICOS:
OBTENÇÃO DE CONHECIMENTO DE PADRÕES CELULARES PARA
OTIMIZAÇÃO DE PROCESSOS BIOTECNOLÓGICOS

Ana Reis de Figueiredo

Março/2010

Orientadores: Nelson Francisco Favila Ebecken
Gilberto Carvalho Pereira

Programa: Engenharia Civil

O mercado mundial de proteínas recombinantes tem se mostrado de grande valor. Porém, em mercados competitivos como a indústria farmacêutica, é de extrema importância a otimização de bioprocessos. Neste contexto, considerando que qualquer produto final será obrigatoriamente sintetizado em alguma fase do ciclo de vida das células, o objetivo principal deste trabalho é demonstrar o potencial da tecnologia de citometria de fluxo no monitoramento de biorreatores acoplada ao desenvolvimento de modelos de redes neurais artificiais envolvidas no reconhecimento de padrões celulares. A taxa de crescimento e atividade metabólica da bactéria *Escherichia coli* DH10b foi estabelecida em culturas do tipo *batch*. O emprego do fluorocromo SYBR Green I permitiu acessar a distribuição do conteúdo de DNA ao longo das diferentes fases do ciclo celular e estabelecer as diferenças fisiológicas na população bacteriana. Em outra abordagem, foi demonstrado o nível de viabilidade celular através da utilização de CFDA. Um modelo de rede neural do tipo Multilayer Perceptron treinado como o algoritmo BFGS e otimizado por um algoritmo genético apresentou alto grau de reconhecimento da heterogeneidade celular ao longo do cultivo. Concluiu-se então, que o acoplamento destas duas abordagens apresenta grande potencial na otimização de biorreatores.

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

CYTOMETRIC DATA MINING:
KNOWLEDGE DISCOVERING OF CELLULAR PATTERNS FOR OPTIMIZATION OF
BIOTECHNOLOGICAL PROCESS

Ana Reis de Figueiredo

March/2010

Advisors: Nelson Francisco Favilla Ebecken
Gilberto Carvalho Pereira.

Department: Civil Engineering

The worldwide market for recombinant proteins has proved of great value. However, in very competitive markets is of utmost importance to optimize bioprocess. In this context, considering that any final product will be compulsory synthesized at some stage of the life cycle of cells, the main objective of this work is to demonstrate the potential of the technology of scanning flow cytometry for monitoring bioreactors coupled with the development of models of artificial neural networks involved Pattern Recognition. The growth rate and metabolic activity of the bacterium *Escherichia coli* DH10b was established in batch-type cultures. The use of fluorochrome SYBR Green I enabled to access the DNA content and its distribution along the different phases of the cell cycle and the physiological differences in bacterial population. In another approach, we demonstrated the level of cell viability through the use of CFDA. A Multilayer Perceptron model trained by the BFGS algorithm and optimized through a genetic algorithm had a high performance to the recognition of cellular heterogeneity along the cultivation. It was concluded that the coupling of these two approaches has great potential in the optimization of bioreactors.

Sumário	
Ficha Catalográfica	iii
Epígrafe	iv
Dedicatória	v
Agradecimentos	vi
Resumo	vii
Abstract	viii
Sumário	ix
Glossário	xi

CAPÍTULO I

Introdução	1
1.2 - Objetivos	6
1.3 - Relevância e Contribuição	7
1.4 - Estado da Arte	8
1.5 - Organização do Trabalho	12

CAPÍTULO II

Descrição do Modelo Biológico	13
2.1 – Diversidade bacteriana e ocorrência	13
2.2 – Estrutura e morfologia básica	16
2.3 – Ciclo celular procariótico	19
2.4 - Viabilidade celular	21
2.5 – Curva de crescimento	22
2.6 – Fontes de heterogeneidade	24
2.6.1. – Heterogeneidade genética	24
2.6.2. – Heterogeneidade bioquímica ou metabólica	25
2.6.3 – Heterogeneidade fisiológica	25
2.6.4 – Heterogeneidade comportamental	25
2.7. – Reconhecimento e processamento de sinais (Quorum sensing)	26
2.8 – Redes regulatórias	27
2.9 – Regulação da expressão gênica	31
2.10 – A cepa <i>E. coli</i> DH10b	33

CAPÍTULO III	
Metodologia	36
3.1 - Condições de cultivo	36
3.2 – Tempo de geração e taxa de crescimento	37
3.3 – Protocolo experimental de ciclo celular	38
3.3.1 – Fixação da suspensão bacteriana	39
3.3.2 – Marcação com SYBR Green I	39
3.4 – Protocolo experimental de viabilidade	40
3.4.1 – Controle de células mortas	40
3.4.2 – Controle de células vivas	41
3.4.3 Análise das amostras	41
3.5 – Citometria de fluxo	42
3.6 – Desenvolvimento de modelos	50
3.6.1 – Modelos Estatísticos	50
3.6.2 – Redes Neurais	53
3.6.3 – Algoritmos Genéticos	61
CAPÍTULO IV	
Análise dos Resultados e Discussão	65
CAPÍTULO VI	
Conclusões	84
Referências Bibliográficas	87

Glossário

Algoritmos - Procedimento sistemático (etapa-por-etapa) para solucionar um grupo de problemas.

Algoritmos Genéticos - Classe particular de algoritmos evolutivos que usam técnicas inspiradas pela biologia evolutiva como hereditariedade, mutação, seleção natural e recombinação (ou *crossing over*).

Batch Culture (cultura em batelada) – Sistema de cultura em larga escala, onde as células crescem sob condições específicas com volume fixo de nutriente, temperatura, pressão e aeração controlados.

Bioprocesso – Processo de produção de insumos por microrganismos.

Biotecnologia – Conjunto de técnicas aplicáveis às atividades que associam a complexidade dos organismos e seus derivados, conciliadas às constantes inovações tecnológicas

Ciclos Biogeoquímicos - O movimento dos elementos e compostos essenciais à vida nos vários compartimentos ambientais.

Ciclo Celular - Conjunto de processos que ocorrem em uma célula entre duas divisões celulares.

Citometria de Fluxo – Técnica utilizada para contar, examinar e classificar partículas microscópicas suspensas em meio líquido, em fluxo.

Comunidade – Da ecologia, significa um conjunto de diferentes populações.

Cromossomo - É uma longa sequência de DNA que contém vários genes e outras sequências de nucleotídeos.

Deriva Genética - É um mecanismo que, atuando em consonância com a seleção natural e modifica as características das espécies ao longo do tempo.

Entropia - A entropia é uma grandeza termodinâmica, geralmente associada ao grau de desordem.

Estado Fisiológico – Estado em que a célula se encontra no decorrer do ciclo celular.

Estado Metabólico – Conjunto de reações bioquímicas que ocorrem na célula.

Expressão Gênica - Refere-se ao processo em que a informação codificada por um determinado gene é decodificada em uma proteína

Fagos – Tipo de vírus que infecta à célula bacteriana, onde realiza ciclo lítico ou lisogênico (bacteriófago).

Fase Capsular – Quando a bactéria apresenta cápsula ou envoltório.

Fase Flagelar – Quando a bactéria apresenta flagelo com atividade.

Fed-batch culture (cultura do tipo fed-batch) – É um tipo de cultura celular contínua.

Fenótipo - São as características visíveis de um indivíduo, que são definidas pela expressão do seu genótipo (isto é, do seu patrimônio hereditário), somada à influência exercida pelo meio ambiente

Fermentação – É um processo anaeróbio de transformação de uma substância (carboidratos) em outra, produzida a partir de microrganismos, tais como fungos e bactérias.

Fissão Binária – Processo de reprodução assexuada dos organismos unicelulares que consiste na divisão de uma célula em duas, cada uma com o mesmo genoma da “célula-mãe”.

Genótipo – Conjunto global de genes de um organismo.

mRNA – Uma molécula de RNA mensageiro.

Mutação – Mudanças na sequência dos nucleotídeos do material genético de um organismo.

Nicho Ecológico – Posição trófica específica de um organismo.

Operon – Vários genes agrupados em uma molécula de mRNA.

Parede Celular – Envoltório externo à membrana citoplasmática das bactérias.

Peritríquio (flagelo) – Distribuição do flagelo em torno de todo o corpo celular.

Plasmídeo – São moléculas circulares de DNA que estão separadas do DNA cromossômico.

População – Conjunto de indivíduos da mesma espécie.

Redes Neurais – São sistemas não lineares que imitam o mecanismo de processamento do cérebro humano.

Sistemas Biológicos – Sistemas que ocorrem no interior de células, como respiração, fotossíntese, etc.

Stress Ambiental – Conjunto de fatores ambientais (externos) que exercem efeitos sobre a célula.

Transcrição – processo pelo qual a informação genética do DNA é repassado para o mRNA

Transposons – Pequeno elemento genético (DNA) móvel.

Viabilidade Celular – Células com capacidade de divisão celular.

CAPÍTULO I

Introdução

A tecnologia de bioprocessos é atualmente empregada na produção de várias mercadorias economicamente importantes, tais como, produtos de química fina, enzimas e proteínas recombinantes terapeuticamente ativas. Como indicativo de importância mundial do mercado na produção de proteínas recombinantes, tendo como por exemplo os biofarmacológicos, cita-se o aumento do volume mundial de 13 bilhões de euros, no ano de 2000. O monitoramento e o controle dos bioprocessos representam um desafio sempre crescente na área da engenharia devido à necessidade econômica, a natureza complexa do crescimento microbiano e a formação de produto em cultivos do tipo *batch* e *fed-batch*. Para conseguir exploração ótima da produção de um organismo em particular, o avanço no melhoramento da capacidade de monitoração e controle dos bioprocessos, situou-se como pivô para a redução dos custos de produção, aumento da quantidade do produto e a manutenção da qualidade (CLEMENTSCHITSCH & BAYER, 2006).

De acordo com JENZSCH *et al.* (2006) os produtos biológicos são conhecidos como estruturalmente complexos. Aparentemente, pequenas mudanças no processo de manufatura podem causar significantes diferenças em suas propriedades clínicas. O uso de microrganismos para produção de moléculas complexas de interesse terapêutico requer o controle preciso de numerosos fatores como concentrações de substrato extracelular e condições da cultura, para regular a atividade dos microrganismos e otimizar o processo (DIAZ *et al.*, 1995). Para BHOWMIK *et al.* (2000) a fermentação é um bioprocessos muito importante para a produção de fármacos e cultura de linhagens celulares. Os substratos usados nestes processos oneram a produção, o que torna o monitoramento em tempo real essencial para a indústria.

Em mercados competitivos, tais como o farmacêutico, a busca constante pela eficiência do processo é essencial. Sensores permitem o monitoramento e controle do progresso de bioprocessos em termos do ambiente físico-químico (pH, temperatura, pressão, oxigênio dissolvido) e processos dinâmicos (taxa de alimentação, taxa de

consumo de oxigênio). Entretanto, parâmetros biológicos como concentração de biomassa, morfologia, estado metabólico, heterogeneidade da população e qualidade do produto não estão diretamente disponíveis, ou somente presentes de forma *off-line*, o que acarreta atraso no tempo de resposta (GLASSEY *et al.*, 1997). Para CLEMENTSCHITSCH & BAYER (2006), as análises que fornecem informações sobre a evolução do bioprocessos, quando realizadas de maneira *on-line* terão benefícios em relação aos métodos de monitoração *at-line* e *off-line*, uma vez que permitirão obter informação do processo em tempo real, o que torna possível a tomada de decisão como estratégia do processo.

Os fermentadores contínuos requerem sistemas de controle apropriados. Segundo BUNTEMAYER *et al.* (1994) nos últimos dez anos, a demanda especial por células cultiváveis causou o desenvolvimento de *setups* complexos nestes equipamentos. KONG *et al.* (1998) por sua vez citaram que uma atenção especial deveria ser colocada na procura dos requerimentos nutricionais de altas densidades de células que surgem de processos contínuos. Diferentemente dos métodos convencionais, que fixam o *setpoint* de parâmetros físicos, o controle de processos contínuos envolve a manipulação de um ou mais fluxos de alimentação do biorreator para controlar a entrada de nutrientes, sendo assim mais sofisticados e complexos. Devido a esta complexidade, um sistema computacional de controle integrado torna-se indispensável em tais aplicações. A manutenção dos nutrientes em um intervalo desejado reduz a geração de subprodutos e aumenta a eficiência na utilização dos mesmos. Estratégias de controle variam a partir da regulação do nível de um único nutriente chave, ou vários fluxos de alimentação de nutrientes balanceados.

A maioria dos sistemas de reatores biotecnológicos tem características extremamente não lineares e uma dinâmica variável ao longo do tempo, portanto difíceis de modelar. Segundo BARBU *et al.* (2005) tais problemas mais difíceis são:

- A ausência de medidas precisas na capacidade de reprodução dos experimentos, tornando difícil o procedimento de identificação e também a escolha da estrutura dos modelos;
- A variação dos parâmetros no tempo, levando em consideração um conjunto de fases distintas da evolução dos bioprocessos;

- A ausência de sensores confiáveis para medir as variáveis do processo com o propósito principal de conhecer o funcionamento interno do bioprocesso.

Durante um processo biotecnológico é necessário monitorar a proliferação de células, bem como sua viabilidade. Isso pode ser reconhecido pela formação colônias em meios sólidos (UFC-unidades formadoras de colônias), habilidade em responder à adição de nutrientes no meio, presença de potencial de membrana, integridade da membrana e atividade metabólica. De acordo com PORTER *et al.*, 1997 a característica obtida em contagens microscópicas mostra uma suave tendência dos dados, enquanto o uso da técnica de citometria de fluxo apresenta grandes variações servindo para enfatizar o controle de qualidade.

As informações a respeito da concentração de biomassa são importantes na tomada de decisão, durante o controle do processo. Estas informações podem permitir a escolha do produto na concentração adequada, de forma a obter-se a produtividade máxima, ou ainda, a ativação de sistemas indutíveis no tempo correto, de modo a otimizar a eficiência global do processo. A existência de uma fração significativa de células permeabilizadas, consideradas mortas ou dormentes, durante qualquer parte da evolução do bioprocessos reflete-se negativamente no rendimento global, uma vez que não contribuem para a formação do produto. Portanto, é importante a obtenção de informação rigorosa e precisa sobre o estado fisiológico das células. Logo, a monitoração da concentração de produtos, da biomassa, dos nutrientes e de outros substratos deve ser também acompanhada da monitoração do desenvolvimento das subpopulações em uma cultura microbiana, no decorrer do processo.

O principal objetivo do controle de bioprocessos é atingir grande eficiência, o que depende do metabolismo celular. Se as células desenvolvem-se normalmente, sob ótimas condições ambientais então, a eficiência do processo será boa. Esta é a razão pela qual muitos pesquisadores têm procurado a possibilidade de determinar uma medida do metabolismo celular (o estado fisiológico das células).

Olhando para o bioprocessos como uma caixa preta, muitas das aproximações para otimização são empíricas. Algumas razões para falhas no controle de bioprocessos são os atrasos substanciais na criação de novos sensores e a descoberta de importantes variáveis-chave para obter conhecimento sobre o estado metabólico. A capacidade de medir condições intracelulares usando ferramentas genômicas, bioinformática ou a tecnologia de citometria ótica em tempo real pode vir a facilitar a engenharia de

biosistemas. Verifica-se que os modelos matemáticos quando utilizados para prever a evolução da biomassa ao longo de um processo biológico de fermentação são muitas vezes imprecisos, pois se baseiam no pré-suposto de que uma população de bactérias é homogênea do ponto de vista fisiológico, o que não é verdade. Desta maneira, os métodos de controle devem ser capazes de lidar com este fluxo de dados, a fim de serem implementados como sistemas de comunicação computacional em grandes plantas operacionais (MANDENIUS, 2004). A mudança de modelos clássicos para técnicas e projetos avançados se mantiveram latentes, devido à alta complexidade. Segundo SIMON & KARIM, 2002, por exemplo, mudanças adicionais incluíram um número exato de sensores necessários para promover medidas adequadas, o conhecimento da dinâmica da planta operacional e suas constantes. Pode-se concluir portanto, que o campo de monitoramento, modelagem e controle de bioprocessos estão em constante desenvolvimento. A Figura 1.1 demonstra a necessidade de integração nas diferentes fases dos bioprocessos.

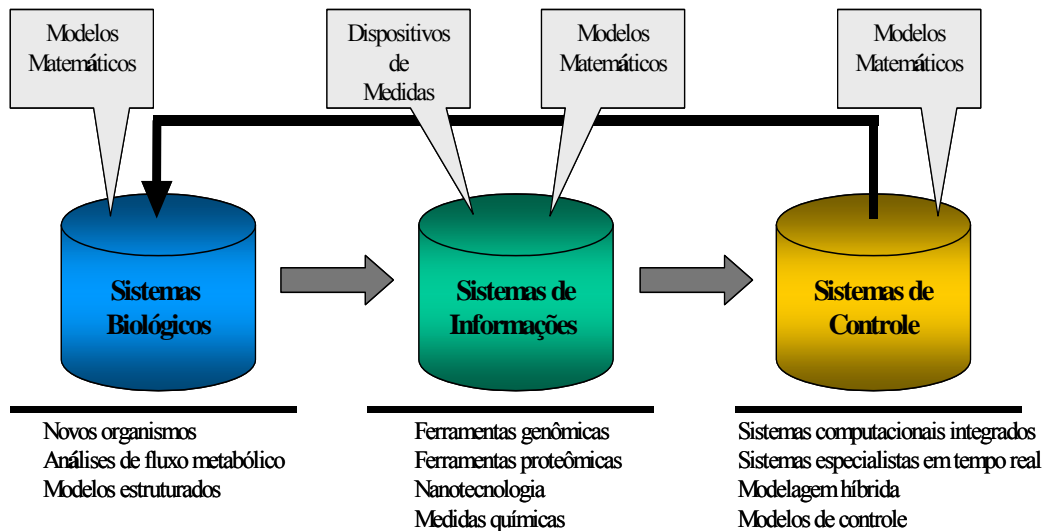


Figura 1.1 – Proposta de integração dos bioprocessos. Fonte: Modificado de MANDENIUS, 2004.

Os sistemas biológicos são naturalmente complexos e necessitam de uma abordagem metodológica nova, que possa modelar os dados relacionados aos bioprocessos. Neste âmbito, muitas vezes modelos puramente matemáticos não são suficientes por serem dados não lineares. Segundo LIAO *et al.* (2003), a aplicação de

técnicas de mineração de dados tem sido bem sucedida em diversas áreas, inclusive em sistemas biológicos.

De acordo com FAYYAD *et al.* (1996) e HAN & KAMBER (2001) a Mineração de Dados (MD) é considerada a principal fase do processo de Descoberta de Conhecimento em Grandes Massas de Dados (*Knowledge Discovery in Large Database - KDD*) que pode ser visualizada de maneira esquemática na Figura 1.2. Esta fase é exclusivamente responsável pelo algoritmo minerador, ou seja, o algoritmo que diante da tarefa especificada, busca extrair o conhecimento implícito e potencialmente útil nos dados. A mineração de dados é, na verdade, uma descoberta eficiente de informações (padrões) válidas e não óbvias de uma grande coleção de dados. A proposta de extrair conhecimento de bases de dados surgiu devido à capacidade crescente no seu armazenamento em meios magnéticos e a necessidade de aproveitá-los.

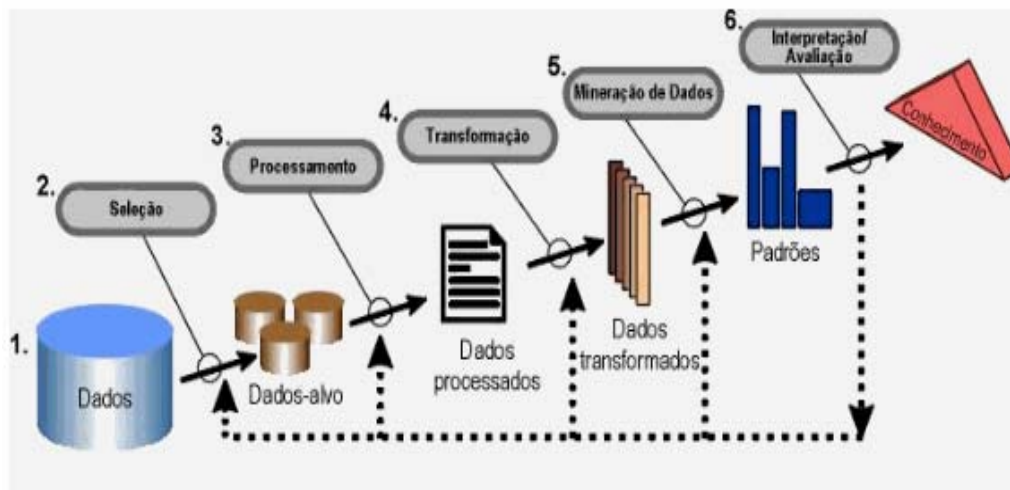


Figura 1.2 – Visão geral das etapas que compõem o KDD. Fonte: FAYYAD *et al.*(1996).

Outro fator que contribuiu em muito para o aumento do interesse em mineração de dados foi o desenvolvimento das técnicas de aprendizado de máquinas (*machine learning*), redes neurais artificiais, algoritmos genéticos, entre outras, que tornaram a descoberta de relações interessantes em bases de dados mais atrativas. Quando se fala de mineração de dados não se considera apenas as consultas complexas e elaboradas que visam ratificar uma hipótese gerada pelo usuário em função dos

relacionamentos existentes entre os dados, mas a descoberta de novos fatos, regularidades, restrições, padrões e relacionamentos.

Além dos algoritmos, a mineração de dados envolve diversas áreas e técnicas. Na Figura 1.3 são demonstradas as etapas da mineração de dados, onde as caixas representam áreas e técnicas.

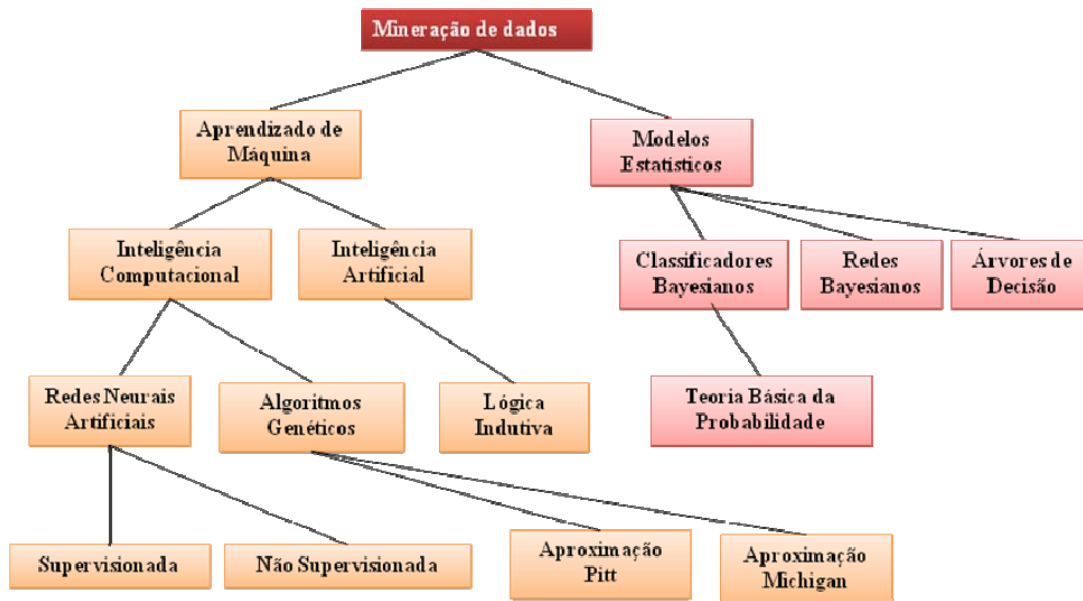


Figura 3 – Etapas da mineração dos dados (*data mining*). Fonte: FAYYAD *et al.*(1996)

1.2 – Objetivos

Segundo SILVA *et al.* (2004) a necessidade de se obter informação no decorrer de bioprocessos, tem contribuído para o aparecimento e utilização de uma grande diversidade de técnicas e ferramentas para este fim. O dado obtido não só tem permitido aprofundar o conhecimento dos processos como também o desenvolvimento e aplicação de novas técnicas. Durante os últimos dez anos, os modelos de redes neurais surgiram como uma ferramenta para modelagem de processos não lineares (KARIM & RIVERS, 1992; SIMON & KARIM, 2002). Por outro lado, a citometria de fluxo surgiu como uma técnica consistente e confiável possibilitando a caracterização individual de células em uma população (RAJWA *et al.*, 2008). Esta técnica pode ser empregada, por

exemplo, na identificação do conteúdo do DNA e RNA, estudos de fases do ciclo celular, estudo dos cromossomos, viabilidade e atividades metabólicas, dentre outras. Portanto, os principais objetivos deste trabalho são:

- 1) Monitorar o processo de divisão celular bacteriana através da tecnologia de citometria de fluxo focada a laser;
- 2) Testar a nova tecnologia do equipamento CytoSence como ferramenta de monitoramento de culturas celulares;
- 3) Estabelecer a curva de crescimento da cultura de *E coli* DH10b através da técnica de citometria de fluxo e compará-la com medidas tradicionais de densidade ótica obtidas por espectrofotometria;
- 4) Estabelecer a porcentagem de células viáveis na cultura;
- 5) Obter uma base de dados citométricos com as informações necessárias para desenvolver modelos de redes neurais artificiais capazes de reconhecer as várias fases do processo de divisão celular bacteriana e seus “estados fisiológicos”.

1.3 – Relevância e Contribuição

O objetivo da otimização da produção de proteínas recombinantes em bactérias é otimizar a exploração do potencial de síntese das células hospedeiras. Para aumentar a produção e permitir qualidade consistente o objetivo dos fermentadores industriais está na otimização de processos e na possibilidade de subida de escala. Devendo manter as condições de reação de modo homogêneo e ótimo para minimizar o estresse microbiano e, ao mesmo tempo, aumentar a precisão metabólica. O modo como as bactérias regulam a progressão do ciclo celular a nível molecular é um problema fundamental e pouco entendido. Atualmente têm-se verificado pesquisas que demonstram circuitos genéticos integrados e altamente complexos (VOHRADSKY, 2001; ARMITAGE *et al.*, 2003; KIKUCHI *et al.*, 2003; MARTINEZ & COLLADOVIDES, 2003; McADAMS & SHAPIRO, 2003; YING *et al.*, 2007).

Até o presente momento, não foi encontrado na literatura a aplicação do modelo de redes neurais artificiais em dados de citometria de fluxo com o objetivo de estudar as fases do ciclo de divisão celular em procariotos. Neste sentido, a principal contribuição deste trabalho é a aplicação de métodos estatísticos e o desenvolvimento de um modelo de redes neurais artificiais que, monitorando a evolução de uma cultura de

E. coli, acopla as suas diferentes fases de ciclo celular com o conteúdo de DNA, que expressa a condição fisiológica das células. Em outras palavras, assume-se que o conteúdo de DNA expresso nas diferentes fases do ciclo celular pode ser usado como medida do estado fisiológico de uma célula. Outra contribuição importante é que este modelo será gerado a partir dos dados citométricos, permitindo mapear em tempo real as condições biológicas da cultura.

1.4 – Estado da Arte

A biotecnologia utiliza microrganismos vivos para a manufatura de produtos muito valiosos. O recente desenvolvimento das técnicas de manipulação genética, o notável avanço no projeto de biorreatores e nos processos de controle baseados em métodos computacionais fundaram uma “nova tecnologia”, que aumentou as possibilidades técnicas. Verificou-se um grande desenvolvimento, tanto nas áreas de projetos dos biorreatores, no desenvolvimento de novos sensores, nos métodos computacionais adaptativos, além de forte investigação para ter acesso às condições fisiológicas dos organismos em cultivo, com o objetivo de automação e aumento de performance dos sistemas biológicos. Portanto, nesta parte do trabalho tenta-se apresentar uma revisão bibliográfica dos esforços nas diferentes áreas que integram a nova tecnologia. Chama-se a atenção para o fato que estas diferentes áreas do conhecimento evoluíram independentemente.

Segundo ASSIS & MACIEL FILHO (2000), deve-se notar que o desenho do sistema de controle de biorreatores não deve ser rígido por causa de:

- Incerteza significativa nos modelos;
- Falta de sensores on-line confiáveis que podem detectar importantes variáveis de estado;
- Natureza não linear e altamente variável dos sistemas;
- Baixa resposta do processo, em particular para células e concentrações metabólicas.

NEBE-VON-CARON *et al.* (2000) atestaram que a montagem da citometria de fluxo *on-line* em unidades de fermentação garantiria um aumento na produtividade

do processo. A estratégia de controle do processo seria baseada na adição de uma fonte de energia, que seja dependente do número de células com um baixo potencial de membrana, permitindo reduzir o número de células que não contribuísem para a síntese do produto desejado, e, por conseguinte, aumentar a produtividade global do processo.

Neste sentido ANDREW & BAILEY (2000), investigaram a estrutura da comunidade bacteriana e seus respectivos estados fisiológicos em um biorreator industrial do tipo Vitox usado como sistema de biorremediação. Os autores demonstraram, através de hibridização *in situ*, uma população heterogênea composta de grupos de *Proteobacterias*, *Citophaga-Flavobacteria* e *Pseudomonas*. Por outro lado, na área computacional, SILVA *et al.* (2000) apresentaram um modelo neural híbrido do tipo *feedforward* no processo de produção de cefalosporina C. O modelo integra a rede neural e equações de balanço de massa em medidas *on-line* da biomassa, substrato disponível e concentração de produto.

HEWITT *et al.* (2001) trabalhando com altas densidades de células em um fermentador do tipo *feed-batch* com a cepa W3110 de *Escherichia coli* usaram citometria de fluxo e múltiplos fluorocromos como estratégia para investigar o efeito da limitação de glicose no estado fisiológico de células individuais. Já PIGRAM & McDONALD (2001) usaram um modelo de redes neurais do tipo Multilayer Perceptron (MLP) para prever a qualidade do efluente de um biorreator industrial de biorremediação na área da refinaria da Baía de São Francisco. Concomitantemente, CHO *et al.* (2001) desenvolveram uma estratégia de controle automático em tempo real baseado em modelos de redes neurais para reatores do tipo *batch* que operam em fluxo contínuo para processos de desnitrificação e desfosforização de efluentes. De outra forma verificou-se que HEWITT & VON-CARON (2001) efetuaram uma aplicação industrial de um citômetro de fluxo multi-paramétrico (*Coulter Epics Elite Analyser*) com o objetivo de acessar o estado fisiológico de células microbianas de *E. coli*, *Rhodococcus* sp. e *Saccharomyces cerevisiae* em fermentadores. Os autores usaram a mesma estratégia de múltiplos fluorocromos anteriormente testada (Rhodamina 123/ iodeto de propídeo, bis-oxonol/ iodeto de propídeo, ou bis-oxonol/brometo de etídeo/ iodeto de propídeo). A aplicação da técnica de citometria de fluxo permitiu ainda neste ano, identificar a distribuição de células eucariotas durante as várias fases do ciclo celular. Neste sentido, NUNEZ (2001) mediu as diferenças no conteúdo de DNA nas quatro fases do ciclo celular de *Saccharomyces cerevisiae*.

Em outra abordagem, NA *et al.* (2002) introduziram os algoritmos genéticos com o objetivo de otimização adaptativa em culturas de leveduras recombinantes em biorreatores do tipo *feed-batch*. Com o avanço das pesquisas sobre a expressão gênica e a aplicação de tecnologia de *microarray* foi possível acessar o comportamento de milhares de genes, conseqüentemente foram produzidas quantidades enormes de dados, que necessitavam de processamento eficiente. Neste sentido, BICCIATO *et al.* (2002) descreveram a aplicação de um modelo de rede neural auto-associativo capaz de classificar e identificar padrões de expressão gênica e marcadores de fenótipos específicos e seus respectivos estados fisiológicos, enquanto SIMON & KARIM (2002) implementaram uma rede neural para expressar a apoptose como função da viabilidade celular, já que tal função não é diretamente disponível.

CIMANDER *et al.* (2003) integraram 1800 sinais provenientes de diversos equipamentos (espectrômetro de massa, espectroscopia infravermelha, probes eletroquímicos, fluorímetros e múltiplos sensores de detecção de gases) em uma plataforma de um sistema especialista operando em tempo real para efetuar diversas tarefas computacionais: regressão do mínimo-quadrado parcial, análise de componentes principais, tomadas de decisão heurística e controle adaptativo de um bioprocessamento. Já POPOVA *et al.* (2003) demonstraram a capacidade de um modelo de rede neural para estimativa da concentração da biomassa e de taxas específicas de crescimento para controlar cultivos de *S. cerevisiae*, enquanto FETECAU *et al.* (2003) utilizaram uma rede neural do tipo *feedforward* para estimar a idade média (*mean age*) como indicadora do estado fisiológico de microrganismos concluindo que o modelo funciona bem mesmo quando o conjunto de dados de treinamento apresenta “ruído branco”. Paralelamente ABU-ABSI *et al.* (2003) apresentaram o desenho de uma nova cubeta de citômetro de fluxo que permite a aplicação de fluorocromos e reagentes em tempo real, favorecendo assim o uso destes novos citômetros na concepção de novas “drogas” e outras aplicações, que requerem *high-throughput screening*.

Os processos de morte celular descritos na literatura são a necrose e a apoptose de células de mamíferos. O primeiro é um processo passivo que ocorre quando as células são expostas a condições ambientais extremas e estresse fisiológico. Porém, na apoptose a célula é induzida a cometer “suicídio” sob condições fisiológicas normais. Alguns autores já sugeriram que a morte celular programada não está limitada aos eucariontes, mas também pode ser ativada em células procariotas como o *Streptococcus*

pneumoniae e *Aeromonas* sp (NING *et al.*, 2002; ENGELBERG-KULKA *et al.*, 2003; RICE & BAYLES, 2003). BREHM-STECHER & JOHNSON (2004) sugeriram que a técnica citométrica pode ser também utilizada como ferramenta para elucidar este processo em bactérias.

ROCA *et al.* (2004) desenvolveram um sistema especialista baseado em conhecimento para supervisionar e controlar um fermentador utilizado na produção de etanol através de cultivos de células de *S. cerevisiae* envolvendo medidas precisas de temperatura, pH, taxa de fluxo e produção de CO₂.

Em 2005 VINTERBO *et al.* citaram que diversos algoritmos de Mineração de Dados tais como *Support Vector Machines*, redes neurais e regressão logística têm sido aplicados na classificação de dados de expressão gênica e que comumente produz-se modelos de difícil interpretação para biólogos e biomédicos. Se regras simples e precisas fossem induzidas a partir de pequenas amostras de dados de treinamento a interpretação dos modelos seria grandemente facilitada, portanto, os autores apresentaram um modelo de regras *fuzzy* para a interpretação da expressão gênica. Ainda neste ano, LOOSER *et al.* (2005) detectaram através de citometria fluxo mudanças no estado fisiológico de *E.coli* que expressavam uma proteína heteróloga de membrana em cultivos do tipo *feed-batch* sob condições de limitação de carbono

Em 2006 verificou-se que HRISTOVA & PATARINSKA compararam a performance de um modelo de rede neural híbrida com outro *feedforward time delay neural network* em cultivos contínuos, com o objetivo de analisar o efeito de memória.

Em 2007 BUSAM *et al.* utilizaram uma abordagem de modelo de rede neural artificial (*single-layer perceptron*) na tarefa de agrupar (*clustering*) dados de citometria de fluxo em bactérias sensíveis ao arsênico e cultivadas em biorreatores. Concomitantemente PATNAIK (2007) comparou um modelo de rede neural e outro cibernético para controlar a síntese de poli-β-hidroxibutirato por *Ralstonia eutropha*. Por outro lado, TEGEL *et al.* 2007 fizeram uma análise dos efeitos de promotores na solubilidade da expressão de proteínas recombinantes baseada em citometria de fluxo. Com o uso de um algoritmo baseado na máxima entropia de deconvolução e a expressão dos dados por uma cultura microbiana altamente sincronizada ROWICKA *et al.* (2007) puderam obter picos da regulação da expressão de transcrição gênica no ciclo celular.

Portanto, é evidente que qualquer tentativa para a produção de moléculas complexas através de bioprocessos, a substância em questão será sempre sintetizada em alguma etapa do ciclo celular do microrganismo cultivado. Desta maneira, todos os modelos ou sistemas que se propõe ao controle de bioprocessos deve considerar e reconhecer estas diferentes fases celulares. Até o presente momento não se verificou a aplicação de um modelo de redes neurais em dados de citometria de bactérias para identificação das fases do ciclo celular sem o uso de marcadores celulares específicos, que são onerosos e demandam citômetros com cubetas mais especializadas.

1.5 – Organização do trabalho

O presente capítulo contextualizou as diferentes faces do tema abordado, a contribuição e os objetivos a serem atingidos, além de uma revisão da literatura. Os capítulos seguintes apresentam-se como:

O capítulo seguinte descreve o modelo biológico usado neste trabalho, visando dar uma idéia da complexidade envolvida na otimização de bioprocessos bacterianos.

O capítulo III é dedicado a apresentar a metodologia e as técnicas utilizadas.

No capítulo IV apresentam-se e discutem-se os resultados obtidos.

No capítulo V encontram-se as conclusões e sugestões para trabalhos futuros.

Por fim, são apresentadas as referências que contribuíram para a realização deste trabalho.

CAPÍTULO II

Descrição do Modelo Biológico

Apesar da complexidade e variedade dos seres vivos todas as células podem ser classificadas em procariotas e eucariotas, com base em sua estrutura e microscopia. Os seres procariontes formam um grupo heterogêneo de microrganismos que incluem as eubactérias (bactérias e cianobactérias) e as arqueobactérias. Neste capítulo descreve-se de forma resumida a morfologia básica das bactérias, seu processo de divisão celular, o conhecimento atual de como estes organismos percebem, interagem e são influenciados pelo ambiente em que se encontram, além de explicar o modelo biológico utilizado. Tais noções são fundamentais para entender a complexidade envolvida nas ações relativas à otimização de bioprocessos.

2.1 Diversidade bacteriana e ocorrência

A diversidade microbiana pode ser verificada em relação a variação de formas e tamanhos, as estratégias metabólicas, a motilidade, aos mecanismos de divisão celular, a biologia do desenvolvimento, à adaptação aos ambientes extremos, composição química, necessidades nutricionais e a fonte de energia.

As bactérias existem a mais de três bilhões de anos e representam provavelmente a forma de vida mais antiga, abundante e diversa na biosfera (Figura 2.1). Segundo SOGIN *et al.* 2006 estes organismos, na natureza, são responsáveis por grande parte do processamento do carbono orgânico e apresentam um importante papel como mediadores em todos os ciclos biogeoquímicos. Encontram-se distribuídos em todos os nichos do planeta, inclusive associados ao homem como parte de sua microbiota (COSTELLO *et al.*, 2009). Algumas espécies podem estar associadas à processos infecciosos, como oportunistas ou como patógenos obrigatórios. De acordo com MÜLLER (2007) não existem instrumentos que avaliem com segurança sua atividade na comunidade ou em relação a outras células. Seu grande potencial evolutivo

pode ser verificado nas diferentes formas de vida que refletem seus respectivos nichos ecológicos, tipos de vida, metabolismo, composição da parede celular e hospedeiros específicos (DOOLITTLE, 1999). A variedade de nichos procarióticos pode ser observada ao considerar as estratégias empregadas na produção de energia metabólica. Alguns procariotos como as bactérias de cor púrpura convertem a energia luminosa em energia metabólica, sem produzir oxigênio. Já as cianobactérias produzem oxigênio na presença de luz.

Nos habitats naturais as bactérias dificilmente vivem em colônias isoladas de espécie única, como pode ser visualizado em culturas de laboratório. Muitas vezes podem ser encontradas aderidas as superfícies como componentes de um biofilme microbiano. A ligação de uma célula a uma superfície corresponde ao sinal da expressão de genes específicos da formação do biofilme. Este pode ser encontrado em qualquer parte onde haja água e um suporte sólido (substrato) para ele se desenvolver. Esta propriedade pode causar vários danos para a indústria, devido a obstrução de dutos e estruturas de sustentação, e para a saúde humana, como na formação da placa bacteriana dos dentes.

As bactérias desenvolveram estratégias sofisticadas para se adaptarem aos vários ambientes. Estudos recentes têm demonstrado que o contexto ambiental desenvolveu uma importante função na produção de fenótipos. Um aspecto importante é a adaptação ao estresse ambiental, o que envolve mudanças globais na fisiologia da célula e na expressão gênica. Para BERTONI & DEHÒ (2007) o estudo de cada fenômeno fornece uma contribuição para o conhecimento dos mecanismos regulatórios globais e específicos, apresentando portanto, grande relevância ambiental, médica e implicações biotecnológicas.

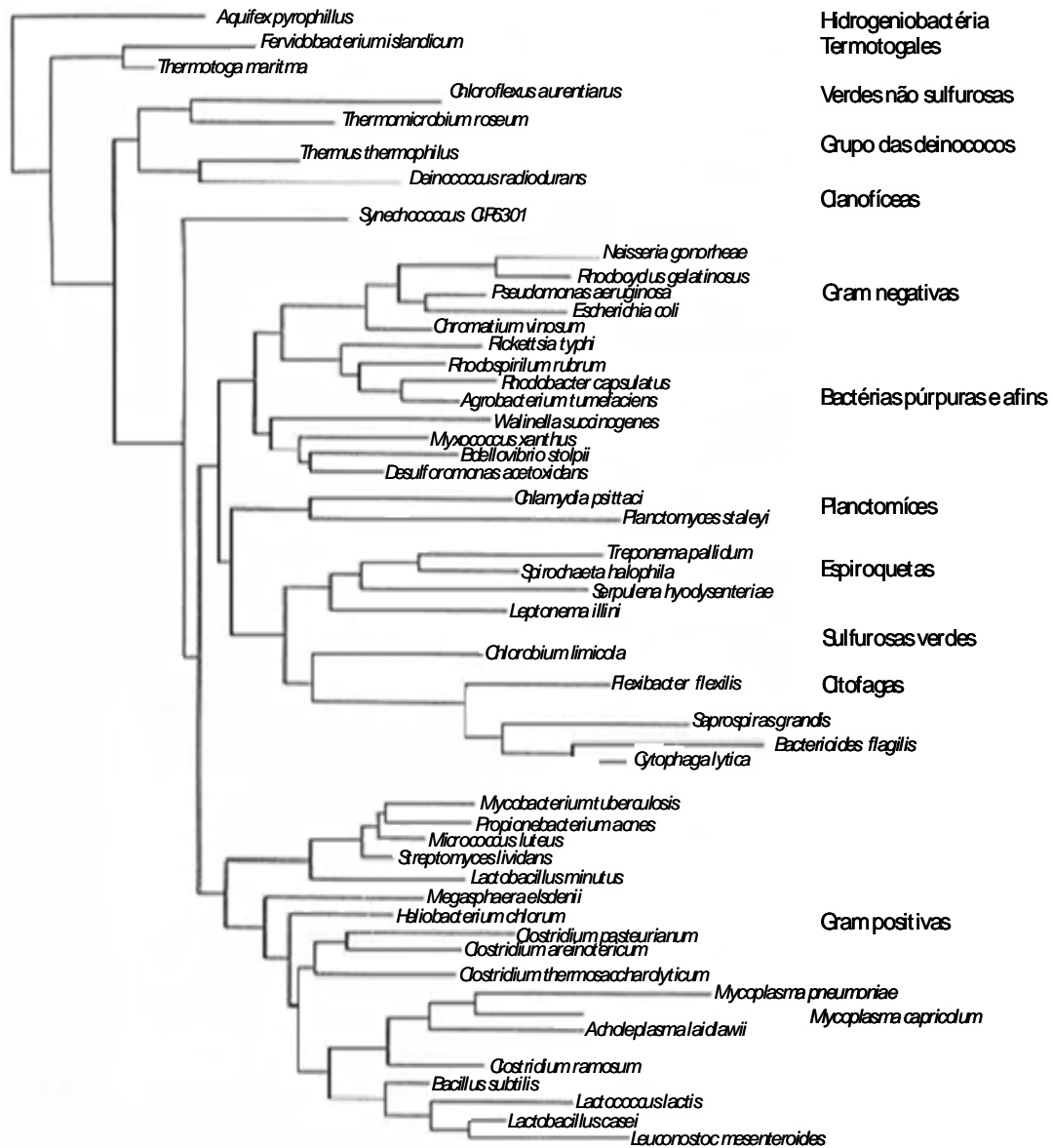


Figura 2.1 - Árvore filogenética da diversidade bacteriana. Fonte: Modificado de DOOLITTLE, 1999.

A diversidade bacteriana pode ser observada na Figura 2.1, onde se encontram os principais grupos bacterianos distribuídos segundo a sua afinidade genética.

2.2 Estrutura e morfologia básica

Existem muitos tamanhos e formas entre as bactérias. A maioria varia de 0.2 a 2.0 μm de diâmetro e de 2 a 8 μm de comprimento e apresenta forma distinta como de cocos (esférico), bacilos (bastão) e espiral. Os cocos normalmente são redondos, mas podem ser ovais, alongados ou achatados em uma das extremidades. Após a divisão celular os cocos podem ser encontrados em vários arranjos como aos pares, ou em forma de cadeias ou em cachos. A maioria dos bacilos encontram-se isolados. A forma de uma bactéria é determinada geneticamente, sendo na maioria das vezes monomórfica. Entretanto, várias condições ambientais pode modificar sua forma.

A célula bacteriana apresenta externamente uma membrana citoplasmática recoberta pela parede celular. A composição química da membrana citoplasmática é na forma de uma bicamada fosfolipídica com proteínas globulares no seu interior, o que facilita a permeabilidade seletiva, o transporte ativo e a respiração celular.

A parede celular bacteriana pode ser identificada pelo método de Gram, que é uma técnica de coloração que permite separar a maioria das bactérias de interesse médico em dois grandes grupos, chamados bactérias Gram positivas e bactérias Gram negativas. As Gram positivas apresentam na composição da parede 90% de sua massa seca constituída por peptidoglicanos, além do ácido teicóico e lipídeos. Já as bactérias Gram negativas não apresentam o ácido teicóico e apenas 10% de peptidoglicano, porém suas paredes são recobertas por uma membrana espessa e externa rica em lipídeos e polissacarídeos, chamada lipopolissacarídeo ou LPS. A parede celular bacteriana confere rigidez, forma e resistência a pressão osmótica, além de apresentar receptores para bacteriófagos e proporcionar a aderência à célula hospedeira.

Algumas bactérias também podem apresentar uma estrutura denominada cápsula, cuja composição varia de polissacarídeos (SPE – substância polimérica extracelular) ou polipeptídeo. Esta estrutura encontra-se externamente à parede celular e pode ser uma camada fina e bem delimitada. Sua estrutura confere resistência à fagocitose, serve como reserva nutritiva, proteção a dessecação e facilita a aderência em superfícies inertes (aderência inespecífica).

A locomoção ocorre em mais de 80% das bactérias devido a presença de flagelo, que é um apêndice longo e sinuoso, dividido em filamento, gancho e corpo

basal. O filamento é rico na proteína flagelina. A estrutura e o arranjo dos flagelos variam de acordo com a espécie e estão relacionados com o ambiente em que cada célula vive. Uma célula bacteriana pode apresentar um único flagelo polar, múltiplos e polares ou vários em uma distribuição chamada peritríquia, ao redor do corpo bacteriano (Figura 2.2). Segundo SOUTOURINA *et al.*, 2001, a presença do flagelo confere vantagem à bactéria, pois propicia o encontro de ambiente mais favorável. Em bactérias Gram negativas podem ser observados apêndices semelhantes a pêlos que são mais curtos, mais retos e mais finos que os flagelos. Estas estruturas, chamadas de fímbrias, quando se encontram distribuídas pelo corpo bacteriano com a função de aderência específica (fator de colonização), ou de pili, que normalmente são mais longos que as fímbrias, havendo apenas um ou dois por célula, com função de aderência e transferência de DNA.

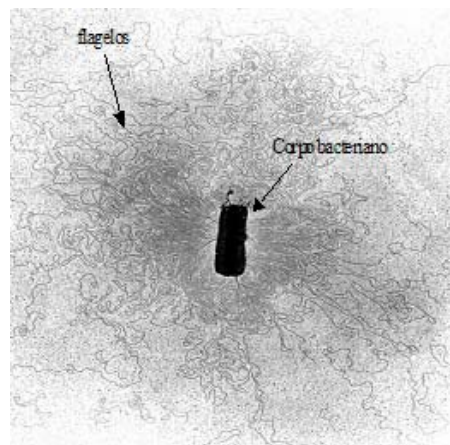


Figura 2.2 – Micrografia de uma célula de *Escherichia coli* com disposição flagelar peritríquia.

Internamente encontram-se os ribossomos, responsáveis pela síntese de proteínas, o mesossomo que é uma invaginação da membrana citoplasmática, cuja função é a síntese e secreção de substâncias e respiração celular. Os grânulos de inclusão, que podem ser metacromáticos ou ricos em polifosfato, polissacarídeos ou ricos em amido ou glicogênio, ricos em enxofre ou em óxido de ferro.

O material genético (DNA) pode ser encontrado na forma de cromossomo (nucleóide), plasmídeo e transposon (Figura 2.3).

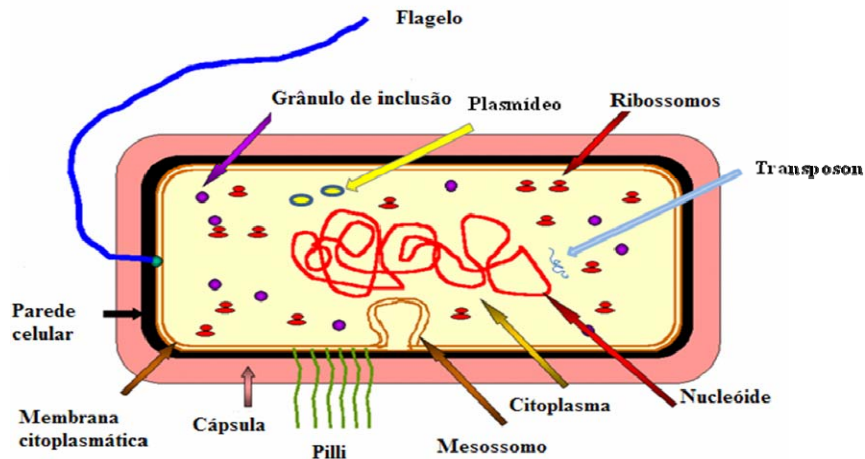


Figura 2.3 – Figura esquemática da estrutura de uma célula bacteriana típica. Fonte: Modificado de <http://pathmicro.med.sc.edu/fox/protobact.jpg>

O DNA da maioria das bactérias é circular com cerca de 1 mm de extensão, representando o cromossomo procarioto, que encontra-se quase sempre único (MADIGAN, 2004). A maior parte dos genes bacterianos é transportada no cromossoma bacteriano, porém genes adicionais são encontrados nos plasmídeos. As unidades de replicação chamadas replicons podem ser encontrados no cromossomo como nos plasmídeos. Porém, nos cromossomos são encontrados genes relacionados ao crescimento, enquanto nos plasmídeos encontram-se genes de atividades especializadas como codificação de resistência a múltiplos antimicrobianos, codificação da pili sexual ou fator F, codificação da síntese de toxinas, codificação de enzimas de catabolismo de açúcares e hidrocarbonetos incomuns e também transmitem genes relacionados a aquisição ou redistribuição do DNA. Os transposons são sequências de DNA menores que incluem informação necessária de um *locus* genético para outro. Os transposons simples transportam apenas sequências de inserção, enquanto os complexos carregam também genes de resistência a antimicrobianos.

2.3 Ciclo celular procariótico

Para um crescimento ótimo, torna-se necessária a coordenação do ciclo celular com os processos metabólicos. A célula bacteriana corresponde a uma usina biossintética, capaz de se duplicar. Os processos sintéticos do crescimento celular bacteriano envolvem várias reações químicas. Algumas destas envolvem transformações de energia, enquanto outras a biossíntese de várias moléculas tais como, co-fatores e co-enzimas necessários às reações enzimáticas. Os processos de crescimento e divisão celular dependem de sequências intrínsecas de eventos de transcrição, que ordenados culminam na duplicação e segregação do cromossomo, seguidos da divisão celular.

Na divisão as células se alongam até atingirem aproximadamente o dobro do seu comprimento, quando então formam uma partição, que eventualmente as separam em duas células filhas. Esta partição é chamada de septo, sendo resultante do crescimento da membrana citoplasmática e da parede celular para o interior da célula em direções opostas até a individualização (Figura 2.4).

Já foram identificadas várias proteínas essenciais na divisão celular de procariotos. Estas proteínas são denominadas proteínas Fts (filamento termosensível). A proteína mais importante deste grupo, a FtsZ, foi extensivamente estudada em *E. coli*. As proteínas Fts interagem e formam um aparelho de divisão denominado divisomo. A formação do divisomo é iniciada pela ligação das moléculas FtsZ, originando um anel ao redor do cilindro celular, localizado na região central da célula, definindo o plano de divisão celular. Acredita-se que o divisomo também contenha proteínas da síntese de peptidoglicanos, entretanto, apesar de ser uma área de pesquisa, acredita-se que o divisomo coordene a síntese da nova membrana citoplasmática e da parede celular em ambas direções, até que a célula atinja o dobro do seu comprimento original. Subsequentemente, forma-se uma constrição, originando duas células filhas (LACKNER *et al.*, 2003).

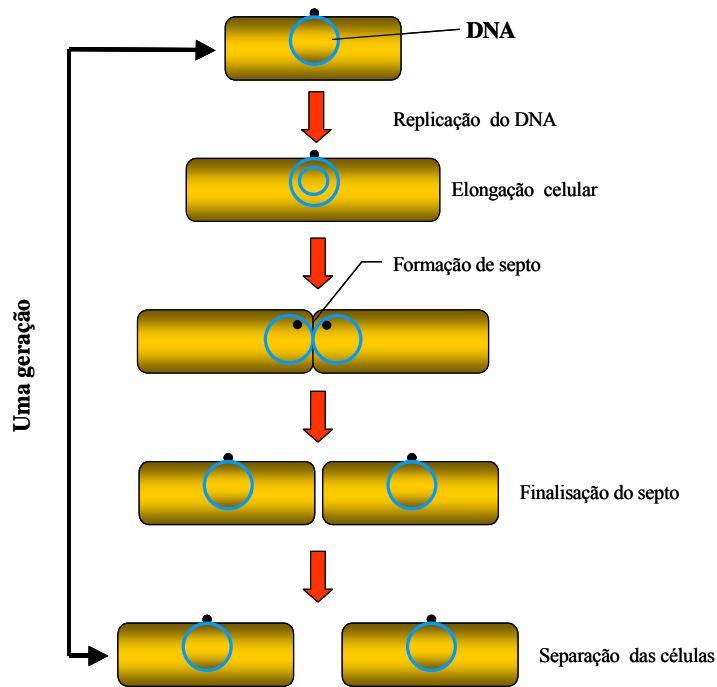


Figura 2.4 - Ciclo de divisão celular simétrico em *E. coli*.

A replicação do DNA ocorre antes da deposição do anel FtsZ. O término da síntese de DNA parece ser o sinal para formação do anel FtsZ, situado entre os nucleóides duplicados. A localização correta da posição central da célula por FtsZ parece ser assistida por uma série de proteínas denominadas Min, especialmente Min E, as quais interagem com os nucleóides duplicados. Em algum momento, a medida que a elongação celular está se processando, as duas cópias dos cromossomos são separadas e encaminham-se para cada uma das células filhas. Simultaneamente à constrição celular, o anel FtsZ começa a sofrer despolarização, promovendo a invaginação dos componentes da parede celular, eventualmente separando as duas células filha. A FtsZ apresenta atividade enzimática, hidrolisando a guanosina-tri-fosfato (GTP) e gerando energia. Acredita-se que esta energia promova a polimerização e despolarização do anel FtsZ (LACKNER *et al.*, 2003; KRSTIC *et al.*, 2007; SENGUPTA & RUTENBERG, 2007).

Até o presente momento, a literatura sobre o ciclo celular dos organismos procariontes foi amplamente e curiosamente desconectada da principal corrente de trabalhos de fisiologia microbiana. Nestes trabalhos foram estudadas culturas contínuas

ou não sincronizadas. Existe um debate entre os especialistas se o acúmulo de biomassa através do ciclo de divisão celular de *E. coli* está mais próximo de ser linear ou exponencial. Os modelos diferem em torno de 6%, no entanto, verifica-se uma vasta diferença biológica (DAVEY & KELL, 1996). Entretanto, os estudos ilustram que a expressão de uma proteína em particular está longe de ser independente das fases do ciclo celular e o conhecimento a este respeito é muito rudimentar. A distribuição de células através do ciclo celular tende a afetar de maneira substancial a performance microbiana e vice-versa.

2.4 Viabilidade

Classicamente a viabilidade em bactérias é definida pela capacidade destas formarem colônias em meios de cultura sólidos ou proliferarem em meios de cultura líquidos. Isto define que as bactérias estão “vivas”, com capacidade de replicação, sendo possível quantificá-las. Porém, este sistema de detecção de viabilidade é dependente do crescimento bacteriano, o que só pode ocorrer na presença dos nutrientes corretos e condições de osmose, temperatura e aerobiose ou anaerobiose. Para os microrganismos fastidiosos ou aqueles que não se adaptam as condições artificiais de laboratório o resultado da viabilidade seria negativo. Tal fato não reflete a realidade das espécies bacterianas, pois existem os microrganismos viáveis mas não cultiváveis em meios. Segundo NEBE-VON-CARON *et. al.* (1999) as contagens que detectam se as células estão saudáveis, injuriadas, em estado de dormência, viável mas não cultivável, assim como quanto as células mortas pode ser obtido de forma direta por métodos óticos.

A técnica de citometria de fluxo associada ao uso de múltiplos fluorocromos é capaz de detectar subpopulações dentro de uma população, o que corresponde a diferentes níveis de funcionalidade da célula. Desta forma pode ser observado em uma população células “viáveis”, mas não cultiváveis, não sendo metabolicamente ativas e nem estando mortas. As células saudáveis são delimitadas por uma membrana seletiva, onde os sistemas de transporte geram um gradiente eletroquímico importante para o funcionamento da célula. Quando a célula encontra-se sob estresse alguns dos sistemas de transporte são afetados e com isto há uma despolarização da membrana. A detecção da atividade metabólica, reveladora do crescimento e associado a divisão celular é

facilmente detectada de forma clássica, mas pode haver metabolismo mesmo na ausência de crescimento, que pode produzir efeitos indesejáveis como degradação, acumulação de toxinas ou a transferência de genes.

O uso de fluorocromos que possam detectar a atividade metabólica celular é uma forma segura de detecção de células metabolicamente ativas.

2.5 Curva de crescimento

O termo crescimento populacional bacteriano refere-se a um aumento do número de células e não ao aumento das dimensões celulares. Quando os microrganismos são cultivados em um sistema do tipo *batch* as concentrações de nutrientes sofrem um declínio enquanto aumentam as concentrações de produtos de síntese e de degradação. O tempo de geração (tempo necessário para que uma célula se duplique) é variável para as diferentes espécies bacterianas, podendo ser de 10 a 20 minutos ou até dias. Este, não corresponde a um parâmetro absoluto, uma vez que é dependente de fatores genéticos, nutricionais e ambientais, indicando o estado fisiológico da cultura.

O crescimento de microrganismos que crescem por fissão binária pode ser representado graficamente como o logaritmo decimal do número de células versus o tempo de incubação. O crescimento de uma população bacteriana é estudado por análise da respectiva curva de crescimento que se caracteriza por quatro fases bem distintas: fase lag, fase logarítmica ou exponencial (log), fase estacionária e fase de declínio ou morte celular (Figura 2.5).

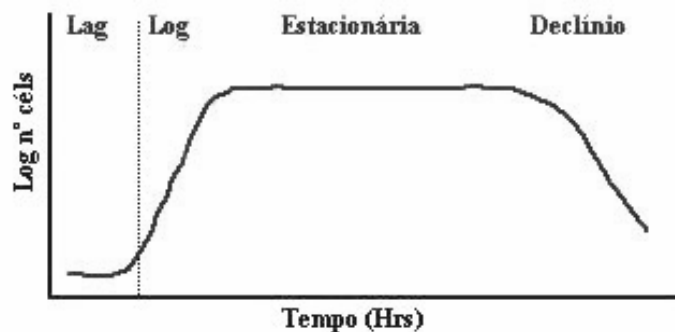


Figura .2.5- Etapas da curva de crescimento bacteriano de uma cultura do tipo *batch*. Fonte: <http://vsites.unb.br/ib/cel/microbiologia/crescimento/crescimento.html>

Fase Lag: Período variável, onde ainda não há um aumento significativo da população. Ao contrário, é um período onde o número de organismos permanece praticamente inalterado. Esta fase é apenas observada quando o inóculo inicial é proveniente de culturas mais antigas. A fase lag ocorre porque as células de fase estacionária encontram-se depletadas de várias coenzimas essenciais e/ou outros constituintes celulares necessários à absorção dos nutrientes presentes no meio. A fase lag também é observada quando as células sofrem traumas físicos (choque térmico, radiações) ou químicos (produtos tóxicos), ou quando são transferidas de um meio rico para outro de composição mais pobre, devido a necessidade de síntese de várias enzimas. Assim, durante este período observa-se um aumento na quantidade de proteínas, no peso seco e no tamanho celular.

Fase Log ou exponencial: Nesta etapa, as células estão plenamente adaptadas, absorvendo os nutrientes, sintetizando seus constituintes, crescendo e se duplicando. Deve ser levado em conta também que neste momento, a quantidade de produtos finais do metabolismo ainda é pequena. A taxa de crescimento exponencial é variável, de acordo com o tempo de geração do organismo em questão. Geralmente, procariotos crescem mais rapidamente que eucariotos. Nesta fase são realizadas as medidas de tempo de geração. Geralmente, ao final da fase log, as bactérias passam a apresentar fenótipos novos, decorrentes do processo de comunicação denominado *quorum sensing*.

Fase Estacionária: Nesta fase, os nutrientes estão escasseando e os produtos tóxicos estão tornando-se mais abundantes. Nesta etapa não há um crescimento líquido da população, ou seja, o número de células que se divide é equivalente ao número de células que morrem. Na fase estacionária são sintetizados vários metabólitos secundários, que incluem antibióticos e algumas enzimas. Nesta etapa ocorre também a esporulação das bactérias. Foram detectados alguns genes (*sur*) que são necessários à sobrevivência das células na fase estacionária. Além destes, existem outros genes (fatores alternativos da RNA polimerase e proteínas protetoras contra danos oxidativos).

Fase de Declínio ou Morte: A maioria das células está em processo de morte, embora outras ainda encontrem-se dividindo. A contagem total permanece relativamente constante, enquanto a de células viáveis cai lentamente. Em alguns casos há a lise celular. Culturas descontínuas tendem a sofrer mutações que podem repercutir na população como um todo. As próprias condições ambientais tendem a promover variações de caráter fenotípico (reversível) nas culturas.

2.6 - Fontes de Heterogeneidade

Mudanças evolutivas ocorrem em um contexto ecológico extremamente complexo. O problema central no processo evolutivo está relacionado a natureza das forças que mantêm a variação entre os indivíduos em suas populações. Assim, três processos têm sido propostos: a seleção, a mutação, e a deriva genética (OCHMAN *et al.*, 2000). Do ponto de vista da ecologia evolutiva, o grande tamanho populacional e o rápido tempo de geração permitem a coincidência das escalas de tempo ecológico e evolutivo. Isto significa que a dinâmica ecológica das mudanças evolutivas pode ser observada em tempo real (RAINEY *et al.*, 2000).

Deixando de lado a mistura das populações naturais e considerando simplesmente uma cultura axênica de laboratório, DAVEY & KEEL (1996) sugeriram que a heterogeneidade microbiana teria como origem três fontes principais: genotípica (através de mutações), fenotípica (via progressão do ciclo celular) e por causa de mudanças ambientais do micro-habitat. Porém, mais recentemente, BREHM-STECHER & JOHNSON (2004) consideraram quatro fontes principais de heterogeneidade: genética, bioquímica ou metabólica, fisiológica e comportamental, as quais serão brevemente consideradas.

2.6.1. - Heterogeneidade genética: a variabilidade genética através de mutação básica é bem entendida para o modelo *E. coli* e está em torno de 10^{-7} por pares de bases na ausência e 10^{-10} por pares de bases na presença de sistemas de reparação pós-replicativa. Mesmo com o valor de 10^{-10} e assumindo que *E. coli* contenha em torno de 2000 genes com 1000 pares de bases cada. Pode-se supor que a cada geração uma mutação terá originado em média 5×10^{-3} da população; após uma geração 0,995 da população seria do “tipo selvagem”. Após n gerações a proporção de organismos do tipo selvagem, seria estatisticamente $0,995^n$. Isto se torna < 0.5 após 140 gerações, ou 5 culturas sequenciais, nas quais as células cresceriam a partir de uma única colônia de 10^9 células. Obviamente, estes números podem ser modificados, mas a conclusão que a variabilidade mutacional sempre contribuirá para a performance fisiológica via seleção permanece.

Genomas microbianos podem ser plásticos, sendo capazes de mudanças substanciais em curtos períodos de tempo. A heterogeneidade genética em

microrganismos individuais pode surgir de maneira randômica, semi-randômica ou através de eventos programados. Os mecanismos da variabilidade genética incluem mutações, eventos de transcrição randômica, fenômenos relacionados com fagos (transdução e lisogenia), replicação cromossomial e amplificação gênica, cópia do número de elementos genéticos móveis, tais como plasmídeo e transposons, variação na fase capsular ou flagelar, e também heterogeneidade genética intracelular tais como aquelas que surgem a partir da transcrição de múltiplos óperons de rRNA em uma única célula. Assimetrias na distribuição do material genético entre células filhas podem ser importantes na direção de processos de diferenciação. Processos relativos ao ciclo de vida celular, incluindo a acumulação de avarias no DNA ou variabilidade na expressão gênica, podem também ser usados para descrever variabilidade genética entre células bacterianas individuais.

2.6.2 - Heterogeneidade bioquímica ou metabólica em uma população é caracterizada por diferenças celulares individuais na composição ou atividade macromolecular e podem surgir de processos fisiológicos relativos ao ciclo celular, tais como *turnover* ou a partir de eventos relacionados com ciclo celular. Assim como a expressão fenotípica do fenômeno genético, a heterogeneidade bioquímica também pode ser originada a partir de mutações e eventos programados associados à diferenciação e transcrição randômica. Da mesma forma que os ácidos nucleicos, as proteínas podem também ser distribuídas assimetricamente entre células mães e filhas. Quantidades de certos componentes macromoleculares tais como carotenóides, carboidratos intracelulares, ou polímeros de estocagem de lipídeos podem também variar individualmente entre as células, contribuindo desta maneira para sua heterogeneidade bioquímica.

2.6.3 - Heterogeneidade fisiológica tem seu ramo primário através da progressão do ciclo celular e descreve diferenças morfológicas entre células, incluindo diferenças no tamanho, forma, características superficiais e internas. Fontes de variação fisiológica em bactérias incluem diferenças no volume celular, forma, variação na densidade, e morfologia do nucleóide. Exemplos mais pronunciados de heterogeneidade fisiológica relativa ao ciclo celular ocorrem em organismos que estejam desempenhando processos de diferenciação. A heterogeneidade fisiológica (e bioquímica) pode também ser dirigida por fatores microambientais agindo de forma localizada na célula.

2.6.4 - Heterogeneidade comportamental é a consequência que pode ser observada na variação célula-célula das características fisiológicas e bioquímicas tais como a

presença, número, estado, ou atividade de componentes quimiostáticos e outras rotas de sinalização. Tal variação tem origem nas mutações genéticas ou a partir de processos estocásticos que afetam também a expressão gênica ou a distribuição sub-celular de vias de componentes chaves. A observação da resposta de células individuais ao nível de estímulos quimiostáticos e fotostáticos representa um meio potencial através do qual a heterogeneidade comportamental pode ser explorada.

Como a célula bacteriana percebe, responde e interage com o meio? Onde as sequências genômicas encontram-se exatamente? Quais são seus pares? Quais são suas funções? Estes são questionamentos que ainda não foram respondidos, mesmo tendo-se o conhecimento de grande número de sequências genômicas. Assim pode-se dizer que bactérias são sistemas complexos, organizados e com especializado nível de controle.

2.7 Reconhecimento e Processamento de Sinais (*QUORUM SENSING*)

As bactérias apresentam adaptações morfológicas e fisiológicas como resposta às mudanças no ambiente. A percepção e processamento de informações químicas do ambiente formam a parte central do controle regulatório para as respostas adaptativas. Muitas bactérias usam um mecanismo de comunicação intercelular denominado *Quorum sensing* (QS) para regular a transcrição dos genes envolvidos em diversos processos fisiológicos, como bioluminescência, transferência de plasmídeos por conjugação e produção de determinantes de virulência.

O sistema QS é usado tanto por bactérias Gram positivas quanto Gram negativas para regular uma variedade de funções fisiológicas. Em todos os casos o QS envolve a produção e detecção de moléculas extracelulares sinalizadoras chamadas autoindutoras ou feromônios, que permitem à bactéria monitorar a densidade de sua própria população celular. Uma das mais conhecidas é a *acyl-homoserine lactone* (HSL) molécula sinalizadora intercelular, que se acumula no meio agindo como agente quimiotático de recrutamento de novas células. A Tabela 2.1 apresenta alguns exemplos de HSL (FUQUA & GREENB, 2002).

Tabela 2.1 – Exemplos de moléculas de *acyl-homoserina lactona quorum sensors*, seus reguladores e funções celulares envolvidas.

Bactéria	Reguladores	Sinal	Função Objetivo
<i>Vibrio fischeri</i>	LuxR-LuxI	3-oxo-C6-HSL	Bioluminescência
	AinR-AinS	C8-HSL	Bioluminescência
<i>Pseudomonas aeruginosa</i>	LasR-LasI	3-oxo-C12-HSL	Virulência e biofilme
	RhIR-RhII	C4-HSL	Virulência
<i>Agrobacterium tumefaciens</i>	TraR-TraI	3-oxo-C8-HSL	Virulência, copias de plasmídeo e transferência por conjugação
<i>Erwinia caratovora</i>	CaR-CarI	3-oxo-C6-HSL	Antibiótico carbapenem
	ExpR		e exoenzimas
<i>Pantoea stewartii</i>	EsaR-EsaI	3-oxo-C6-HSL	Exopolissacarídeos
<i>Rhodococcus sphaeroides</i>	CerR-CerI	A7-C14-HSL	Agregação
<i>Vibrio anguillarum</i>	VanR-VanI	3-oxo-C10-HSL	Ainda não identificada

Fonte: Modificado de FUQUA & GREENB, 2002.

Quando a concentração destas moléculas excede um determinado nível ou *threshold*, uma cascata de transcrição de sinais é iniciada, tais como, a ativação de redes metabólicas hierárquicas envolvidas no comando da expressão gênica e na produção de uma determinada proteína (BASSLER, 1999). Estudos recentes (WITHERS *et al.*, 2001; FUQUA & GREENBERG, 2002) mostraram que o QS modula a comunicação intra e inter específicas. Esta expressão é dependente da densidade bacteriana. A informação dos sinais químicos e físicos adquiridos através de QS reflete o *status* das condições ambientais.

2.8 Redes Regulatórias

O conceito de redes de transcrição bacteriana implica que as bactérias podem aprender e se adaptarem às mudanças ambientais, adquirindo inclusive uma “capacidade de memória”. Isto pode ocorrer através da percepção e processamento integrado de uma enorme quantidade de informação, através de receptores na superfície e interior da célula como pode ser visto na Figura 2.6 (WITHERS *et al.*, 2001).

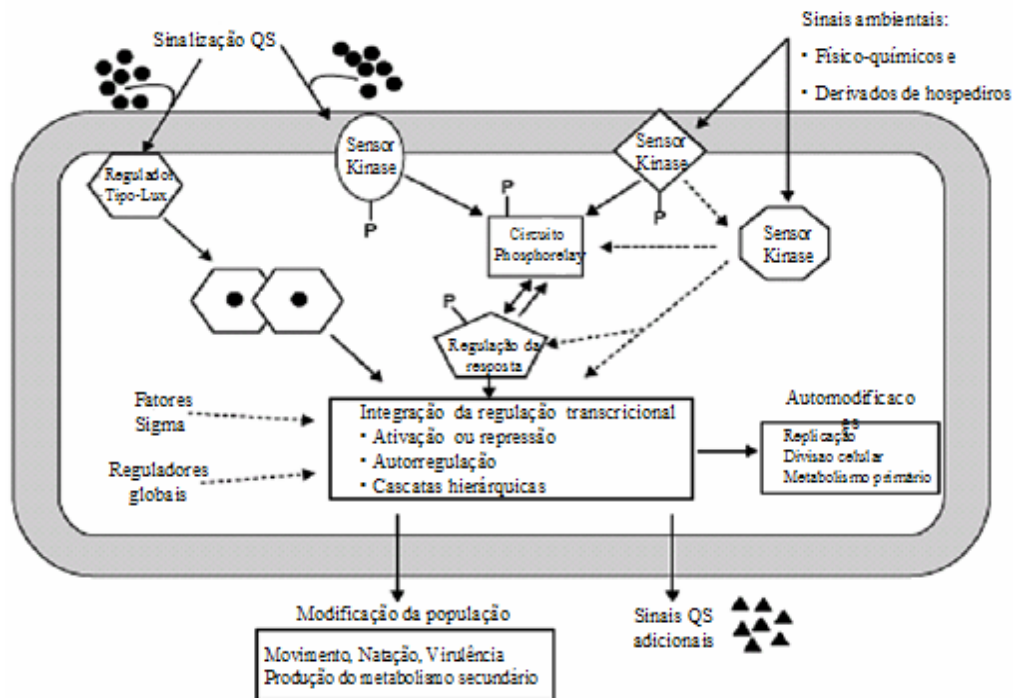


Figura 2.6 – Módulo de integração da sinalização intra e inter QS para adaptação ambiental. Fonte: Modificado de WITHERS *et al.*, 2001.

Diante de funções complexas, tais como, o controle do ciclo celular e processos de adaptação envolvendo centenas de genes amplamente distribuídos pelo genoma, as células microbianas envolvem uma grande variedade de redes reguladoras (Figura 2.7). Os nós destas redes são as unidades de transcrição (genes e óperons) juntamente com seus produtos protéicos, enquanto que, as ligações que os conectam correspondem às interações da regulação transcricional mediada pelo fator de transcrição das proteínas (BALÁZSI *et al.*, 2005).

Evidências recentes indicam que interações potenciais de redes reguladoras de transcrição são usadas diferentemente de acordo com as condições ambientais nas quais a célula existe. Entretanto, as unidades topológicas subordinadas a tal utilização diferencial não são bem conhecidas.

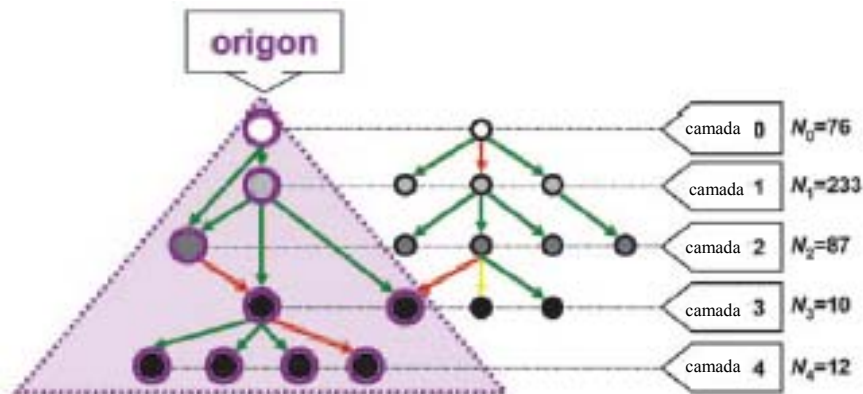


Figura 2.8 – Representação esquemática da rede de transcrição de *E. coli*. Os nós são genes/óperons e seus respectivos produtos. Fonte: Modificado de BALÁZSI *et al.*, 2005.

Estudos recentes sobre a dinâmica das interconexões geraram as primeiras percepções sobre a capacidade de processamento de informação, apesar da posição das recorrências (*building blocks–motifs*) nas redes de regulação transcricional. Sua agregação em grandes estruturas topológicas podem modificar seu comportamento dinâmico (YING *et al.*, 2007). Apesar desses avanços, há uma clara necessidade para decifrar a dinâmica de organização ao nível de sistema da dinâmica de utilização de redes de regulação transcricional engatilhada pelos vários sinais ambientais e intracelulares. Com seus resultados, BALÁZSI *et al.* (2005) sugeriu que *E. coli* usou unidades topológicas específicas da sua rede de regulação de transcrição para detectar componentes elementares (*modes*) de sinais ambientais complexos e subsequentemente desenvolver uma resposta pelo agrupamento destes *modes* elementares próximos a camada de saída da rede.

Em relação a existência dos conceitos de estruturas de regulação multigênica, *origons* são entidades mais complexas que *modulons*, porém menos complexas que *stimulons*. Especificamente, todos os nós em um *modulon* têm de ser controlados diretamente por um *commom*, um regulador pleiotrópico, enquanto que nos *origons* eles podem ser controlados indiretamente por percolação dos níveis transcricionais alterados através da rede. Por outro lado, os *origons* são sub-redes originadas a partir de um único

fator de transcrição, enquanto que *stimulons* incluem todos os nós afetados por um sinal ambiental e são compostos de todos os *origons* enraizados no fator de transcrição sensível para o sinal. Sinais distintos do meio ambiente afetam a seletividade dos *origons*. Baseado na sua complexidade, as perturbações ambientais podem ser classificadas como elementares (mudança de um único fator ambiental) ou complexas (mudanças simultâneas de dois ou mais fatores ambientais). A resposta transcricional da célula dependerá do tipo de perturbação, da estrutura do *origons* afetado e sua interconectividade com outros *origons*. Se uma perturbação elementar é altamente específica para um ou poucos *origons*, isto afetará somente óperons que estão dentro deles. Por outro lado, perturbações complexas envolvendo sinais múltiplos, por exemplo, dois diferentes açúcares, tipicamente afetam duas ou mais proteínas sensoriais dentro do *origons*, resultando em uma resposta combinada de *origons* individuais. Finalmente, perturbações complexas envolvendo estímulos não relacionados, como oxigênio e um açúcar, são freqüentemente processados de forma independente e então combinados por sobreposição e convergência dos *origons* (BALÁZSI *et al.*, 2005).

2.9 Regulação da Expressão Gênica

As células devem ser capazes de controlar a expressão da informação genética para garantir uma coordenação eficiente de numerosas reações químicas, utilizar os recursos disponíveis da melhor forma possível e realizar os processos associados ao seu desenvolvimento. Porém, muitos genes não são regulados, sendo considerados genes constitutivos, ou seja, seus produtos são produzidos constantemente em velocidade fixa. Desta forma algumas enzimas, denominadas constitutivas são necessárias em concentrações semelhantes em todas as fases de crescimento sendo sintetizadas continuamente, enquanto outras, são necessárias apenas sob certas condições. A expressão gênica, por sua vez, é controlada pelos mecanismos de regulação que podem ocorrer durante ou após a síntese da proteína em nível transcricional (regulando a quantidade de mRNA) ou em nível traducional (permitindo ou não a tradução do mRNA em proteína). Vale ressaltar que o número de mecanismos regulatórios é extremamente vasto; e que a maioria dos genes é regulada normalmente por mais de um tipo de mecanismo.

O mecanismo pelo qual as moléculas efetoras afetam a transcrição é indireta, ocorre a partir de sua associação com proteínas regulatórias específicas que afetam a síntese de mRNA. No caso de uma enzima reprimível o co-repressor associa-se a uma proteína repressora específica. Essa proteína repressora alterada pode então ligar-se a uma região específica do DNA, próxima ao promotor do gene denominada região operadora. Esta região originou o termo óperon, correspondente a um conjunto de genes cuja expressão é regulada por um único operador. Todos os genes de um óperon são transcritos como uma única molécula de mRNA. O operador localiza-se adjacente ao promotor onde a transcrição é iniciada. Quando um repressor se liga ao operador, a transcrição é bloqueada, pois a RNA polimerase é impedida de se ligar ou de se mover ao longo do DNA. Assim, a proteína ou as proteínas codificadas por esta molécula de mRNA deixam de ser sintetizadas (WOLF *et al.*, 2001; PRICE *et al.*, 2005a; PRICE *et al.*, 2006). A repressão consiste em um tipo de regulação denominada de controle negativo, a proteína repressora promove a repressão da síntese de mRNA. Entretanto, no chamado controle positivo da transcrição, uma proteína reguladora (ativadora) promove a ligação da RNA polimerase, aumentando a síntese de mRNA.

O mais comum é que as proteínas sejam sintetizadas sob suas formas ativas; essa atividade é subsequentemente reduzida, ou inibida pela ação de compostos celulares específicos geralmente associados às vias metabólicas catalisadas por estas enzimas. Um dos principais mecanismos de controle de ativação enzimática é a inibição por retroalimentação (*feedback*). Tal mecanismo está envolvido, por exemplo, na síntese de aminoácidos ou de purinas onde o produto final inibe a primeira enzima da via biossintética. Se o produto final é consumido, a síntese pode ser restabelecida.

De outra forma, algumas vias biossintéticas são reguladas através de isoenzimas (isofuncionais) que catalisam a mesma reação embora reguladas por mecanismos distintos. Neste caso a atividade enzimática inicial diminui sucessivamente, até atingir valor zero, quando as isoenzimas encontrarem-se em excesso. Outro mecanismo de controle enzimático é a modificação enzimática por intermédio da adição ou deleção de pequenas moléculas orgânicas que promovem alterações conformacionais na enzima.

Embora pequenas moléculas estejam envolvidas com a frequência da regulação da transcrição, raramente atuam de maneira direta. Tais moléculas influenciam a ligação de certas proteínas, denominadas proteínas regulatórias em sítios

específicos do DNA. São na realidade responsáveis pela regulação da transcrição. Dois tipos gerais de interações proteína-ácidos nucléicos são descritos: os não específicos em que a proteína se liga a qualquer região do ácido nucléico e as interações específicas, em que há uma especificidade da ligação com relação a uma seqüência de DNA.

Muitos genes podem apresentar promotores sujeitos a vários tipos de regulação ou possuir mais de um promotor, sendo cada um sujeito a determinado sistema de regulação. Existem conjuntos de mecanismos de controle, sendo a repressão gênica regulada tanto por efetores específicos, associadas à determinada função daquele gene (como resposta de um *regulon*), quanto a efetores mais gerais que respondem a um aspecto mais global do metabolismo. Os organismos precisam regular vários genes diferentes simultaneamente em resposta as alterações ambientais. Os mecanismos regulatórios que respondem a sinais ambientais regulando a expressão de muitos genes são denominados sistemas de controle global (WOLF *et al.*, 2001; PRICE *et al.*, 2005a, PRICE *et al.*, 2005b; PRICE *et al.*, 2006).

2.10 A cepa *E. coli* DH10b

A família Enterobacteriaceae constitui um grande grupo heterogêneo de bacilos Gram negativos cujo habitat natural é o trato intestinal dos seres humanos e de outros animais. Muitos gêneros fazem parte desta família e podem ser móveis com flagelos peritríquios ou imóveis, anaeróbios facultativos, que fermentam a glicose ao invés de oxidá-la, frequentemente com produção de gás, não produzem a enzima oxidase e reduzem o nitrato a nitrito (KONEMAN *et al.*, 1997).

Os membros do gênero *Escherichia* são constituintes do trato intestinal de humanos e outros animais de sangue quente, embora não correspondam aos organismos dominantes nesse habitat. A *E. coli* é uma bactéria Gram negativa, em forma de bacilo, que desempenha papel nutricional importante no trato intestinal, pela síntese de vitamina K e vitaminas do complexo B. Devido a sua natureza aeróbia facultativa, esse organismo provavelmente também auxilia no consumo de oxigênio, tornando o intestino grosso anóxico. Linhagens selvagens de *E. coli* raramente apresentam exigências em relação a qualquer fator de crescimento, sendo capazes de crescer a partir de uma grande variedade de fontes de carbono e energia, como açúcares, aminoácidos, ácidos

orgânicos e outros. Mesmo sendo representantes da flora intestinal normal algumas linhagens de *E. coli* são patogênicas, devido a presença de mecanismos de virulência como produção de toxinas e adesinas. Algumas podem ser enteropatogênicas, estando relacionadas aos quadros de disenteria (TRABULSI *et al.*, 2002).

Muitos fatores favoreceram a utilização da *E. coli* como organismo modelo em estudos de bioquímica, genética e fisiologia bacteriana. Apesar desta bactéria não ter sido a primeira a ter o seu cromossomo sequenciado, continua sendo o principal microrganismo nas pesquisas e aplicações da engenharia genética. O genoma da *E. coli* apresenta cerca de 5 milhões de pares de bases e vários milhares de genes codificando mais de 4000 proteínas, embora cerca de 38% destas proteínas ainda tenham função desconhecida (MADIGAN *et al.*, 2004). A análise genômica inicia-se com a clonagem molecular dos fragmentos do DNA que o compõem, os quais serão posteriormente sequenciados, gerando uma série de informações que podem ser acessadas por meio de banco de dados (*National Center for Biotechnology Information-NCBI*, *DNA DataBank of Japan-DDBJ*, *European Molecular Biology Laboratory-EMBL*). Acredita-se que parte do genoma da *E.coli* foi adquirida por processos de transferência horizontal de genes, isto é, genes originados de outros organismos. O que não resulta em um genoma cada vez maior, pois muitos dos genes adquiridos desta maneira não apresentam vantagens seletivas sendo, portanto perdidos por deleção. A linhagem *E.coli* K-12 tradicionalmente utilizada em estudos de genética obteve parte do seu cromossomo também por transferência horizontal (SANTIAGO, 2005).

O modelo bacteriano utilizado foi a linhagem DH10b de *E. coli* derivada da *E. coli* K-12. O genoma da DH10b foi construído sem o uso das modernas técnicas de biologia molecular, mas através de várias manipulações genéticas. Este genoma apresenta alta eficiência de transformação, habilidade em estabilizar e manter grandes plasmídeos e a falta de sistemas de restrição dependentes de metilação, o que a torna um ótimo modelo biológico para as operações essenciais e diárias que vão desde a propagação de um simples plasmídeo a criação de grandes bibliotecas de clones para a determinação de sequências genômicas inteiras (DURFEE *et al.*, 2008). A Figura 2.9 apresenta o genoma da DH10b, onde os genes que codificam proteínas são mostrados no anel externo em azul, e em laranja encontram-se os genes da fita complementar. As deleções e inserções são indicadas fora do círculo.

A cepa DH10b foi gentilmente cedida pelo Laboratório de Biologia Molecular do Departamento de Bioquímica Médica da Universidade Federal do Rio de Janeiro.

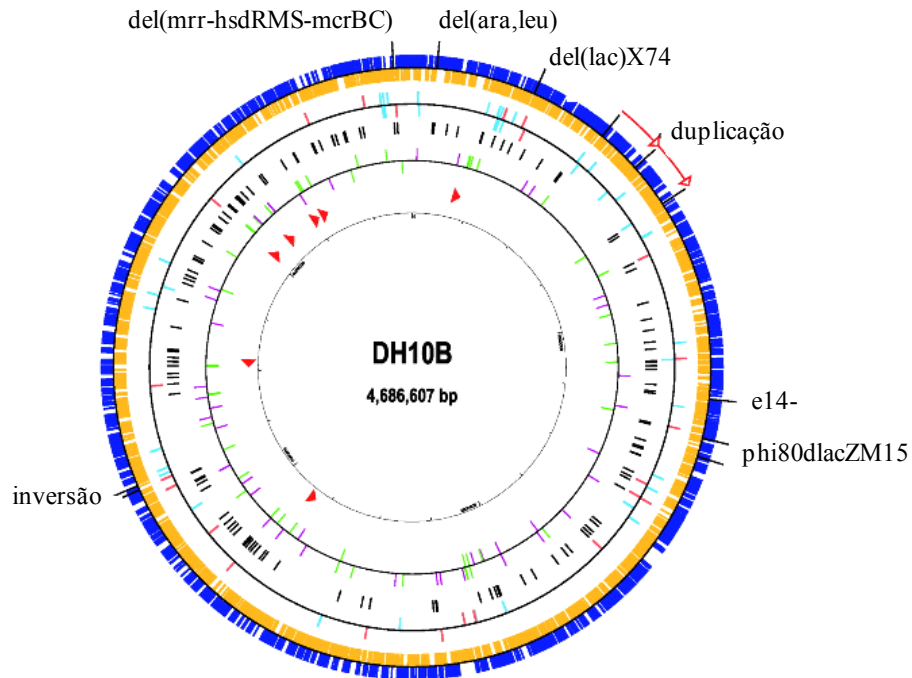


Figura 2.9 Diagrama circular do genoma da *E. coli* DH10b. Fonte: Modificado de DURFEE *et al.*, 2008.

O próximo capítulo apresentará a forma como a cepa DH10b foi cultivada em laboratório, o estabelecimento da curva de crescimento para obtenção dos parâmetros utilizados no ensaio do ciclo celular, o ensaio de viabilidade celular, o protocolo de marcação com fluorocromos o DNA bacteriano e os modelos utilizados para tratamento dos dados.

CAPÍTULO III

Metodologia

Neste capítulo descreve-se as condições de cultivo, a técnica de citometria de fluxo, bem como os métodos estatísticos e técnicas computacionais de aprendizado de máquina utilizado para o reconhecimento de padrões celulares neste trabalho.

3.1 Condições de cultivo

A partir do crescimento da *E coli* DH10b em placa de Petri contendo ágar Luria Bertani (LB Agar - Difco-USA, composição g/L, 10.0g NaCl, 10.0g triptose, 5,0g extrato de levedura, 15g ágar, 1000mL de água destilada), pH 7.0, isolou-se uma colônia que foi inoculada em 1000mL de caldo LB (LB Broth - Difco-USA, composição g/L, 10.0g NaCl, 10.0g triptose, 5,0g extrato de levedura, 1000mL de água destilada), pH 7.0, o qual foi mantido sob agitação constante de 200 rpm à temperatura de 37⁰ C, por 24 horas, para obtenção de massa celular. Em seguida, o crescimento foi fracionado para tubos plásticos do tipo Falcon de 50 mL e centrifugado a 10.000 x g, por 10 minutos. O sobrenadante foi desprezado e a massa celular foi resuspendida em água deionizada estéril e novamente centrifugada. Esta operação foi realizada por 3 vezes consecutivas. Ao final a massa celular obtida foi fracionada em alíquotas de 1,5 mL em uma concentração de 10⁷ células/mL (escala de McFarland), acondicionadas em tubos plásticos do tipo *Eppendorf* e guardadas em geladeira a 4⁰C.

Para a realização da curva de crescimento bacteriano foi realizada uma cultura do tipo *batch* com volume total de 1000mL de caldo LB inoculado com uma alíquota de 1,5 mL da suspensão bacteriana e mantida por 8 horas sob agitação constante de 200 rpm e temperatura de 37⁰ C. O monitoramento da cultura foi realizado através das técnicas de densidade ótica por espectrofotometria (580 nm) e citometria de fluxo focada a laser (488 nm).

Todos os meios de cultura foram reidratados segundo instruções do fabricante, levados a fervura para dissolução completa e esterilizados em autoclave a 121⁰C por 15 minutos.

3.2 Tempo de geração (g) e taxa de crescimento

Como a população se duplica em cada geração, o aumento populacional é dado por 2^n , onde n é o número de gerações. O aumento da população é exponencial ou logarítmica. Estas observações podem ser expressas em forma de equações para o cálculo do tempo de geração.

Se N_0 = Número da população inicial

N_t = Número da população no tempo t

n = Número de gerações no tempo t

$$N_t = N_0 \times 2^n \quad (\text{equação 1})$$

Transformando-se a equação 1 em logaritmo tem-se:

$$\log N_t = \log N_0 + n \log 2 \quad (\text{equação 2})$$

Resolvendo para n tem-se que:

$$n = \frac{\log N_t - \log N_0}{\log 2} = \frac{\log N_t - \log N_0}{0,301} \quad (\text{equação 3})$$

O número de gerações n pode também ser dado expresso por:

$$n = t/g \quad (\text{equação 4})$$

em que

t = tempo de cultura

g = tempo de geração

A taxa de crescimento em um sistema fechado pode ser expressa por meio de uma constante média da taxa de crescimento (K), ou seja o número de gerações por unidade de tempo, muitas vezes expresso em gerações por hora.

$$K = \frac{n}{t} = \frac{\log N_t - \log N_0}{0,301_t}$$

(equação 5)

O tempo de geração médio (g) é o tempo que uma população demora em média para se duplicar, podendo ser calculado. Se a população se duplica então $t = g$ e $n = 1$. Consequentemente, a equação 1 é convertida na sua forma mais simples:

$$N_t = 2 N_0 \quad \text{(equação 6)}$$

Substituindo $2 N_0$ na equação da taxa de crescimento e resolvendo para k tem-se:

$$k = \frac{\log(2N_0) - \log N_0}{0,301_g} = \frac{\log 2 + \log N_0 - \log N_0}{0,301_g} = \frac{1}{g}$$

$$k = \frac{1}{g}$$

(equação 7)

e consequentemente, o tempo de geração

$$g = \frac{1}{k}$$

$$g = \frac{0,301_t}{\log N_t - \log N_0} \quad \text{(equação 8)}$$

3.3 Protocolo experimental de ciclo celular

Devido a inexistência de protocolos de citometria de fluxo específicos para o estudo do ciclo celular bacteriano, realizou-se uma adaptação do protocolo de análise do ciclo celular de leveduras (número 11.13) encontrado no ISAC (2006), cujos os procedimentos são apresentados nas etapas seguintes.

3.3.1 Fixação da Suspensão Bacteriana

O crescimento bacteriano após 120 minutos (fase log de crescimento) foi centrifugada a 10.000 x g, por 10 minutos à temperatura ambiente e o *pellet* “lavado” 2 vezes em água destilada gelada, em 5 vezes o seu volume, e centrifugado a 10.000 x g por 10 minutos. O sobrenadante foi eliminado e o precipitado foi resuspenso em água deionizada gelada, ajustando sua concentração a 1×10^7 cels/mL, seguindo a escala de McFarland.

A suspensão obtida foi fracionada em alíquotas de 1,5 mL, distribuída em tubos de microcentrífuga, e centrifugadas a 10.000 x g, por 10 minutos. O precipitado obtido foi “fixado” com 1 mL de etanol a 10% por 12 horas a temperatura de 4⁰C.

3.3.2. Marcação com Syber Green I

As amostras fixadas foram centrifugadas em microcentrífuga a 10.000 x g por 10 minutos e o precipitado obtido foi “lavado” em 50 mM de tampão citrato de sódio pH 7,5 e centrifugado novamente por 10 minutos. O novo precipitado foi tratado com 250 µL de uma solução 1% de RNase (*DNase free*, Sigma-Aldrich, São Paulo, Brasil - 1mg de RNase em 1 mL de tampão Tris-EDTA, pH 8,0) por 1 hora a 50⁰C, para diminuição do coeficiente de variação das concentrações do DNA bacteriano. Após este tempo foi adicionado 50 µL de proteinase K (Sigma-Aldrich, São Paulo, Brasil - 20 mg de proteinase K em 1 mL de água destilada esterilizada em membrana filtrante de 0,22 µm, com estocagem a 4⁰C) e incubado por 1 hora a 50⁰C, para aumentar a permeabilidade da parede celular facilitando a marcação do DNA.

As suspensões obtidas foram tratadas com 20 µL de Sybr Green I (Sigma-Aldrich, São Paulo, Brasil) solução de trabalho (1µL de Sybr Green I em 10 µL de tampão Tris-EDTA, pH 8,0) à temperatura de 4⁰C, protegido da luz, por 12 horas. Após a marcação fluorescente foi adicionado Triton X-100 0,1% (Sigma-Aldrich, São Paulo, Brasil - 50µL de Triton X-100 em 49,95 mL de tampão PBS 0,1% BSA, pH7.3) até a concentração final de 0,25% e depois foi homogeneizado com o uso de um agitador do tipo vortex. As amostras foram colocadas em banho de gelo e depois sonicadas com 3 pulsos consecutivos de 30 W por 3 segundos, com intervalo de 2 segundos entre os pulsos, para eliminação dos grumos de células, evitando a ocorrência de picos de DNA no momento da leitura. A aquisição de dados relativos ao conteúdo de DNA em células individuais foi efetuada por citometria de fluxo.

A solução tampão Tris-EDTA foi preparada utilizando-se 3,7g de Tris base (*tris hidroxymethyl-aminomethane*) e 0,121g de dissódio EDTA em 900mL de água destilada. Após este preparo o pH foi corrigido para pH 8,0 com 1N de HCl e o volume completado para 1000mL. O tampão foi filtrado em membrana filtrante de 0,22 μm e estocado a 4⁰C por 1 mês.

A solução tampão citrato de sódio foi preparada utilizando-se 14,7g de citrato de sódio em 900mL de água destilada e o pH foi corrigido para 7,5 com 1N de HCl ou 1N de NaOH. Após a correção do pH a solução foi filtrada em membrana de 0,22 μm e estocado a 4⁰C por 1 mês.

O tampão PBS 0,1% BSA foi preparado utilizando-se 0,23g de NaOH₂PO₄, 1,15g de Na₂HPO₄, 9,00g de NaCl, 900mL de água, ajustando-se o pH para 7.3 com o uso de 1N de NaOH ou 1N de HCl. O BSA (soro albumina bovino) foi hidratado em água destilada estéril, segundo recomendação do fabricante, e retirou-se 1 mL que foi misturado ao PBS. Foi adicionado água para completar o volume de 1000mL e filtrou-se em membrana de 0,22 μm , ficando estocado a 4⁰C por 1 mês.

Todos os reagentes utilizados foram comprados na Sigma-Aldrich, São Paulo, Brasil).

3.4 Protocolo experimental de viabilidade

O protocolo de viabilidade utilizado foi o de número 11.3.8 obtido no ISAC (2006), que avalia a viabilidade pela atividade metabólica. As etapas realizadas são descritas abaixo:

3.4.1 Controle de células mortas

Para obtenção do controle de células mortas utilizou-se uma cultura em fase log de crescimento com concentração de 1×10^6 cels/mL, seguindo a escala de McFarland. A cultura foi centrifugada a 12.000 x g por 2 minutos e o precipitado foi ressuspenso em um volum de 990mL de álcool a 70%, ficando em temperatura de 4⁰C por 12 horas. Após este período foi adicionado 10 μL da solução diacetato de carboxifluoresceína (CFDA - Sigma-Aldrich, São Paulo, Brasil) em *dimethyl sulfoxide*

(DMSO - Sigma-Aldrich, São Paulo, Brasil), homogeneizando lentamente, para ser incubado por 30 minutos à temperatura de 37⁰C. O monitoramento de células mortas foi realizado através da técnica de citometria de fluxo focada a laser (488 nm), utilizando-se o sensor FLO (fluorescência laranja).

A solução contendo CFDA foi preparada utilizando-se 25mg de CFDA e 2000µL de DMSO.

3.4.2 Controle de células vivas

Nesta etapa utilizou-se a cultura em fase log de crescimento com concentração de 1×10^6 cels/mL, seguindo a escala de McFarland. A cultura foi centrifugada a 12.000 x g por 2 minutos e o precipitado foi ressuspensado em 990 µL de tampão TE 50mM com pH 7,4 (Tris Cl EDTA). A amostra obtida foi então marcada com 10 µL da solução de CFDA e incubada por 30 minutos à temperatura de 37⁰C.

O tampão TE foi preparado com uma solução 10mM de Tris Cl e uma solução 1mM de EDTA, pH 8,0. O tampão Tris Cl 1M foi preparado dissolvendo-se 121g de de Tris base (tris hydroxymethyl aminomethane) em 800 mL de água. O pH foi ajustado com HCl, aproximadamente 70 mL para o pH 7,4. A solução assim obtida foi misturada sendo acrescida com água destilada até o volume final de 1000mL. Após o preparo, o tampão TE foi autoclavado e guardado em temperatura ambiente.

Todos os reagentes utilizados foram comprados na Sigma-Aldrich, São Paulo, Brasil).

3.4.3 Análise das amostras

Ao longo da curva de crescimento a cultura foi observada em relação à viabilidade pela técnica descrita acima. O procedimento técnico utilizado para as amostras na fase lag e início da fase logarítmica foi retirar 990 µL da cultura e marcar com 10 µL da solução de CFDA, incubando por 30 minutos à temperatura de 37⁰C. Nas fases seguintes do crescimento foram retirados 100mL da cultura para centrifugação como no procedimento realizado para o controle positivo.

3.5 Citometria de fluxo

A citometria de fluxo multiparamétrica é uma técnica importante que pode monitorar, em tempo real, o estado fisiológico de cada célula durante o seu desenvolvimento (CÁNOVAS *et al.*, 2007). Permite desta maneira, uma análise rápida e quantitativa de populações inteiras de células baseada em características individuais. Tipicamente uma taxa de 1000 células/s⁻¹. O equipamento utilizado neste trabalho foi o CytoSub (Figura 3.1) cujo citômetro de fluxo é a tecnologia CytoSense (Cytobuoy *b. v.*, Worden, Netherlands). Este equipamento de última geração efetua análises automatizadas. Sendo portátil e facilmente transportado (DUBELAAR *et al.*, 1999; DUBELAAR & GERRITZEN, 2000), o aparelho é capaz de analisar partículas como as células planctônicas (1 a 1000 µm) e maiores volumes de líquidos (>4 mL).

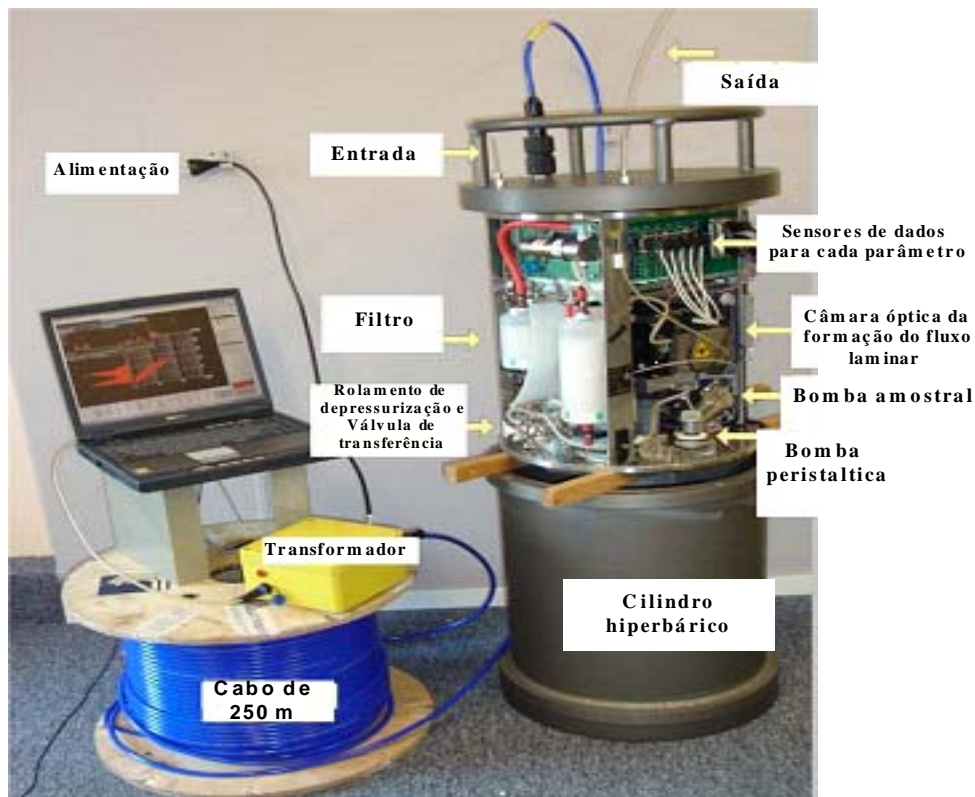


Figura 3.1 – O Citômetro de Fluxo CytoSense em sua versão submersível CytoSub.

O CytoSub é uma tecnologia de Citometria de Fluxo de Escaneamento (*laser scanning cytometry-LSC*) de fase sólida. Os instrumentos baseados nesta tecnologia (LSC) podem fornecer informação visual da morfologia das células e a distribuição espacial da fluorescência dentro de cada célula devido a geração de assinaturas óticas (*pulse shapes*). Tem ainda a capacidade de adquirir múltiplas medidas da mesma célula (Figura 3.2). Geralmente, partículas fluindo ao longo de seu maior eixo (L (μm)), têm a forma de seu sinal do sensor frontal (forward scatter-FWS) definido por $2*(5+L)$ pontos. As medidas do sensor frontal FWS têm sido usadas para estimar o tamanho da célula bacteriana (BOUVIER *et al.*, 2001). Até o presente momento, este é o único citômetro capaz de desempenhar tal tarefa. Segundo SUNRAY *et al.*, 2002 a técnica de citometria de fluxo é adequada para observação de propriedades celulares em função do tempo.

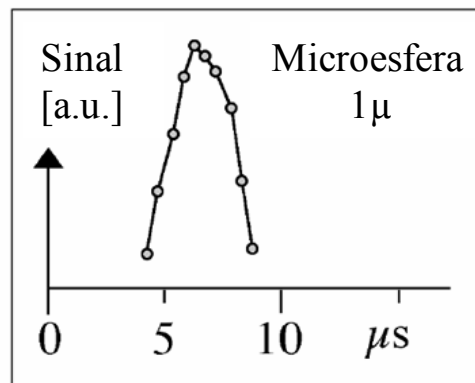


Figura 3.2 – Assinatura ótica de uma microesfera (beads) de 1 micrômetro. Para cada micrômetro a eletrônica do CytoSub é capaz de adquirir até 12 medidas.

A análise individual das partículas se baseia na separação de cada uma, graças a passagem da amostra dentro de um líquido carreador (*sheath fluid*), que devido ao fluxo laminar, evita que haja mistura entre a amostra e este líquido. Isto é chamado de focalização hidrodinâmica. Resumidamente, um tubo de silicone ou de borracha de diâmetro interno de 1 mm (CytoSence) leva a amostra para dentro de uma cubeta de dois estágios (Figura 3.3) onde cada partícula é interceptada por um feixe do laser (5 μm) de comprimento de onda definido. Isto resulta em uma difração da luz e uma difusão acompanhada eventualmente de uma emissão fluorescente endógena, ou exógena quando um marcador fluorescente é excitado. Múltiplas características,

incluindo a contagem, tamanho e conteúdo das células e resposta a probes fluorescentes são usados como diagnóstico da função celular (BREHM-STECHER & JOHNSON, 2004).

No caso do CytoSub, o instrumento está equipado com um laser de safira (Coherent Saphyre, 488 nm, 15 mW). O fluxo da amostra é controlado por uma bomba peristáltica que trabalha em uma velocidade variável de acordo com as necessidades específicas de cada amostra.

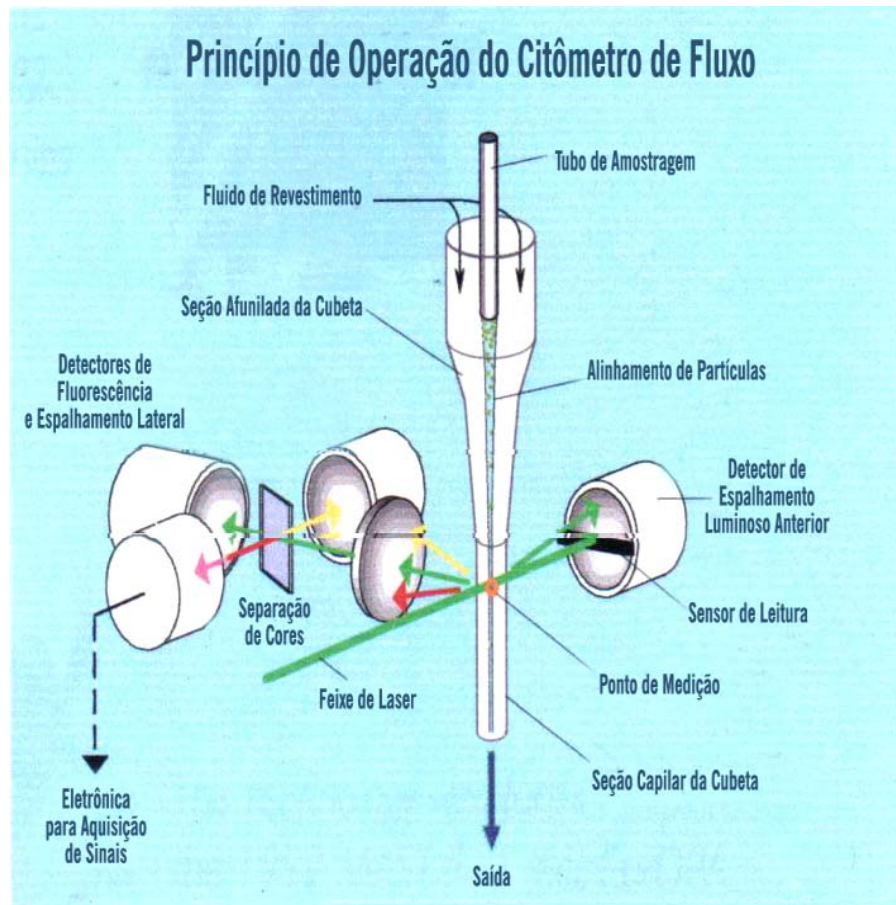


Figura 3.3 - Princípio básico de funcionamento do citômetro de fluxo.

Comumente, o líquido carreador flui a uma velocidade de 2 m s^{-1} e uma vazão de $80 \text{ cm}^3 \cdot \text{min}$. O sensor lateral (*sideward scatter-SWS*) e os sinais fluorescentes são dispersos por um grade holográfica côncava (*concave holographic grating*) e coletados por meio de fotomultiplicadores híbridos (*hybrid photomultiplier- HPMT*). O sinal do sensor frontal FWS é coletado em um fotodiodo (*positive intrinsic negative diode-PIN*), aprimorado (componente semiconductor que permite transformar o sinal luminoso em

sinal elétrico) muito mais sensível que os fotomultiplicadores, pois recebe a maior parte da energia do laser (Figura 3.4). As variáveis SWS e fluorescências são coletadas a 90° de incidência. Um espelho colocado do lado oposto da cubeta permite aumentar cada um dos componentes. O restante da luz emitida é coletado por uma fibra óptica e representa o SWS. A velocidade de digitalização dos sinais correspondem aos perfis das partículas que é de 4MHz. Os pulsos dos fótons são então convertidos em sinais elétricos e depois numéricos antes de ser registrado no disco rígido (64 kb por variável) localizado abaixo da câmara óptica. Somente as partículas com estrutura interna são importantes.

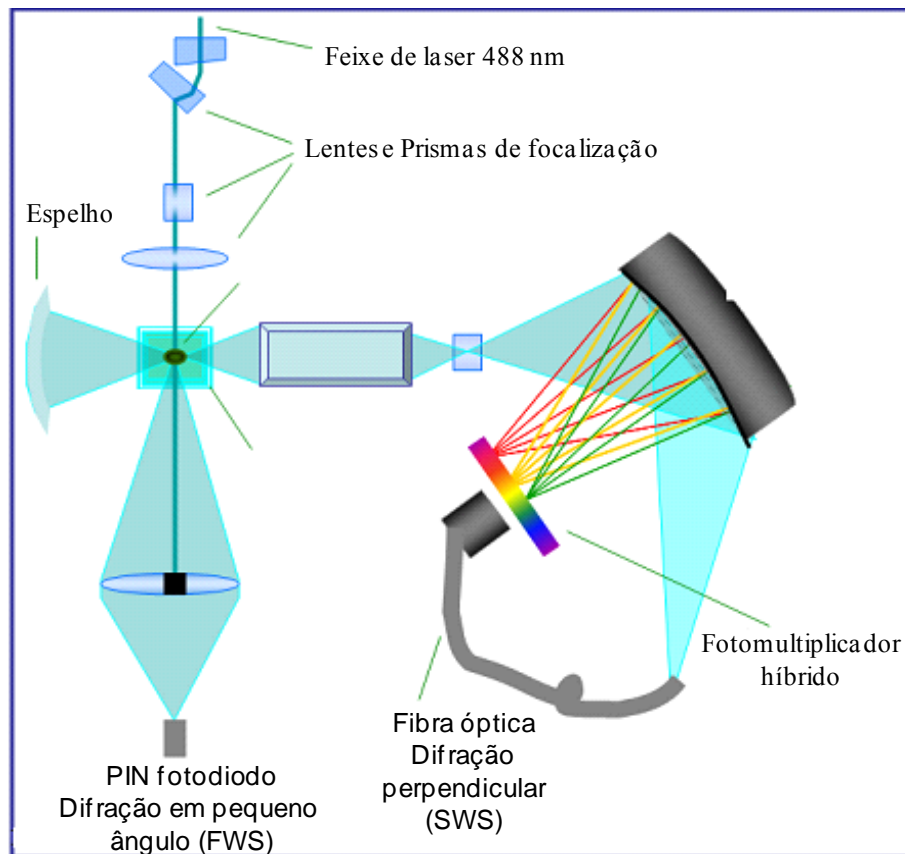


Figura 3.4 – Esquema conceitual da cubeta óptica vista por baixo. Nela as partículas são interceptadas pelo laser e a luz é refratada em pequenos ângulos do sensor FWS dando uma estimativa do tamanho da partícula. O feixe perpendicular é um conjunto importante de sinais de fluorescência que são separados por uma grade holográfica côncava, e a luz do laser difratado, que corresponde a complexidade estrutural da partícula (SWS).

A aquisição de dados é efetuada com o programa CytoUSB fornecido pelo fabricante. A reprodutibilidade das contagens citométricas foram, neste trabalho, determinadas usando-se microesferas de 0,95, 6 e 10 μm (Estapor, Merk -França). Cada frasco contendo estas microesferas foi manualmente homogeneizado, submetido a um banho de ultrassom e analisado 15 vezes com o software CytoClus 3 também fornecido pelo fabricante do equipamento. O coeficiente de variação das contagens no CytoSub (CV_c) foi determinada a partir da média aritmética das contagens (\bar{u}) e seus respectivos desvios padrão (SD).

$$CV_c = \frac{100 * SD}{\bar{u}} \quad (\text{equação 9})$$

A variação da medida de concentração das partículas é a soma da variabilidade do instrumento e da variedade da população. A contagem de partículas raras é significativamente afetada pela Lei de Poisson (SHAPIRO, 2003).

$$CV_p = \frac{100}{\sqrt{n}} \quad (\text{equação 10})$$

O método de calibração das contagens citométricas foi o mesmo utilizado por GAJKOWSKA *et al.*, 2006 que é baseado na determinação de um quadrante (*gates*) com microesferas uniformes, que apresentem a mesma fluorescência e o mesmo padrão de excitação e emissão, como nas amostras a serem analisadas. Desta maneira a contagem dos diversos grupos bacterianos foi efetuado de acordo com a seguinte equação:

$$células / \mu L = \frac{X * Y * Z}{Y} \quad (\text{equação 11})$$

Onde, X é o número de células bacterianas no gate, Y é o número de microesferas em concentração pré-determinada e Z é o fator de diluição, quando aplicado.

Alguns modelos matemáticos simples são atribuídos para cada uma das formas de sinal: total, máximo, média, inércia, centro de gravidade (CG), *fill factor*, assimetria, número de células (DUBELAAR *et al.*, 2003). Todos estes valores são disponibilizados em citogramas que facilitam a identificação dos agrupamentos de células, que apresentem propriedades óticas similares derivadas dos modelos matemáticos (Figura 3.5).

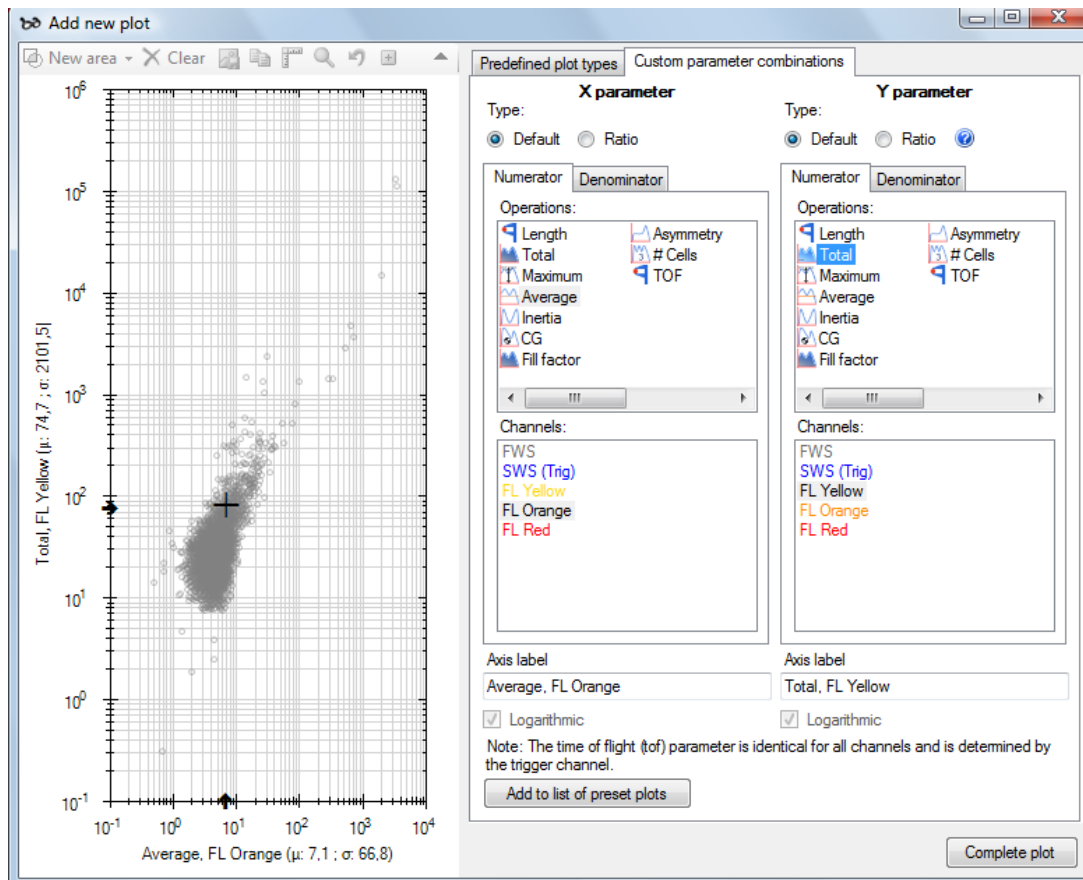


Figura 3.5 – Janela principal do programa CytoClus 3 demonstrando diversos parâmetros de análise de dados

Os modelos empregados nos citogramas são simbolizados e descritos conforme abaixo:

O **Total** - Σ - é simplesmente o valor detectado de cada ponto somado ao comprimento de cada partícula.

O **Máximo** (Max) - é o valor máximo encontrado ao longo do comprimento da partícula através da unidade de detecção.

A **Média** (*Average*) - é igual ao total dividido pelo número de dados, porém removendo-se a dependência do comprimento total da fluorescência.

$$A = \Sigma / N \quad \text{(equação 12)}$$

A **Inércia** (*Inertia*) - é definida como o segundo momento da forma do pulso.

$$I = \sum_0^N (n^2 \cdot D_n) - \frac{CG^2 \Sigma}{\frac{1}{12} \Sigma N^2} = \sum_0^N (n^2 \cdot D_n) - \frac{CG^2}{\frac{1}{12} N^2}$$

(equação 13)

O **Centro de Gravidade** (CG) - é encontrado pela divisão do primeiro momento da forma do pulso pelo total.

$$CG = \frac{\sum_0^N (n \cdot D_n)}{\Sigma}$$

(equação 14)

Fator de Preenchimento (*Fill factor*) - fornece uma indicação sobre a “solidez” do pulso.

$$F = \frac{\sum^2}{N \sum_0^N D_n^2}$$

(equação 15)

A **Assimetria** (*Assymetry*) – indica a distribuição do sinal sobre o comprimento da partícula.

$$As = \left| \frac{2 \cdot CG}{N} - 1 \right|$$

(equação 16)

O **Número de células** (*# cells*) - fornece uma indicação sobre o número de células em uma partícula.

$$NC = \frac{N}{2\pi} \sqrt{\frac{\sum_0^N (D_n - D_{n-1})^2}{\sum_0^N D_n^2 - \frac{\sum^2}{N}}}$$

(equação 17)

O tamanho relativo à intensidade do FWS (*FWS size*) fornece a razão entre o número de células e o número de células pelo FWS.

Dois tipos de valores absolutos correspondem à média da amplitude de cada sinal registrado em seu comprimento. Os valores absolutos de fluorescência, de FWS ou de SWS são expressos sobre a forma da média, mas pode ser recuperado sobre a forma de valor total (correspondendo à integral de cada modelo) por poder comparar a outros instrumentos de medida de fluorescência total.

Os fatores que correspondem à geometria das assinaturas ópticas são obtidos e resumidos na Figura 3.6. Cada partícula é definida por 35 valores discretizados (6 elementos deduzidos das assinaturas * 5 entidades por partícula) acessados pelo programa CytoClus 3.

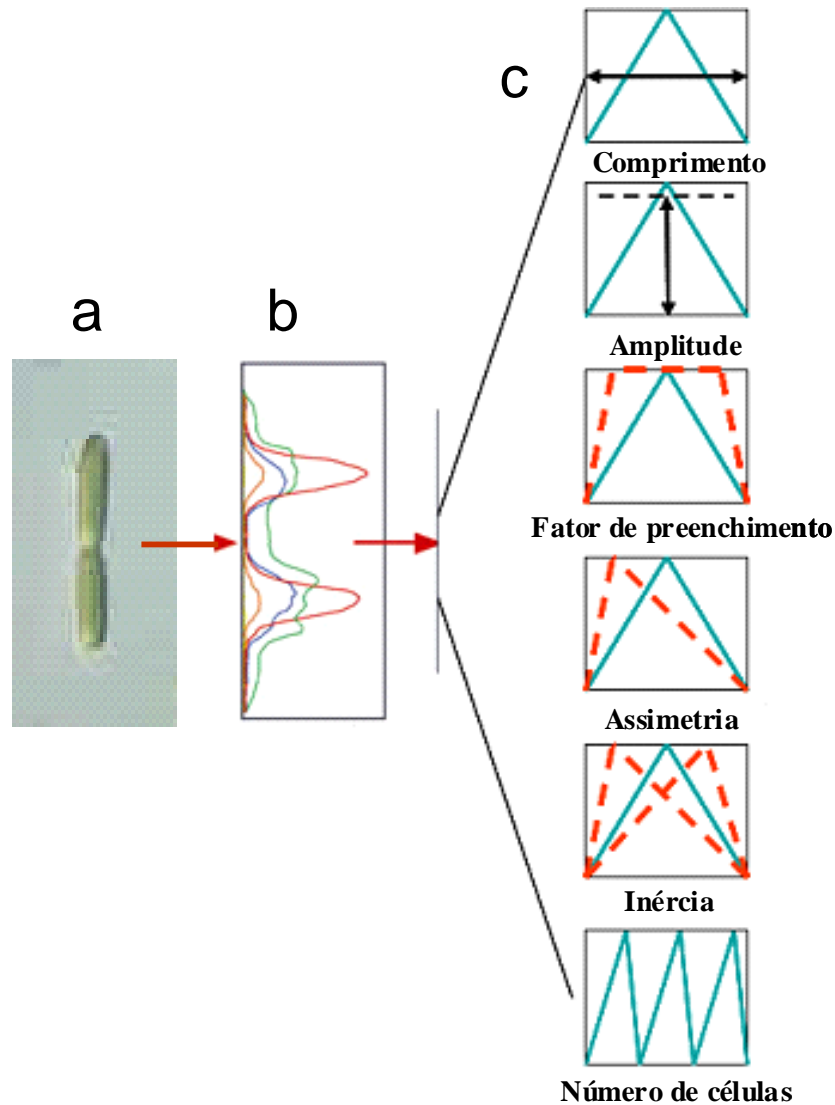


Figura 3.6 - a e b ilustram a sucessão dos procedimentos de um sinal e c, esquematização das variáveis discretizadas obtidas após o registro das assinaturas. Esta discretização permite categorizar as partículas entre si e facilitar seu reagrupamento em elementos similares com o auxílio do CytoClus 3.

3.6 Desenvolvimento de Modelos

Como citado anteriormente os citômetros de fluxo permitem altas taxas de aquisição de dados ($1000 \text{ célula/s}^{-1}$) e de diversos parâmetros (variáveis) que podem ser obtidos (máximo de 47 no caso do CytoSub). Desta maneira a citometria é uma técnica multiparamétrica e a análise dos dados bi ou tridimensionais torna-se extremamente demorada. Contudo, para uma efetiva análise de dados é necessária a aplicação de técnicas computacionais e o uso de modelos que permitam a extração de informação relevante. Essencialmente, o que se tenta fazer nos dados da citometria é executar tarefas de classificação. Portanto, nesse trabalho utilizaram-se abordagens estatísticas e provenientes da área de aprendizado de máquina para este fim.

3.6.1 Modelos Estatísticos

Inicialmente, utilizou-se uma abordagem de estatística multivariada como a técnica de Análise de Componentes Principais (ACP). O objetivo foi verificar uma possível redução de dimensionalidade dos dados a fim de melhorar o desempenho dos algoritmos de identificação e classificação de subpopulações bacterianas. Para a execução desta tarefa foi utilizado o programa *Statistica* 8.0 (StatSoft, Inc.). A ACP é um método de aprendizado não supervisionado uma vez que o agrupamento ocorre através da variância total, sem qualquer tentativa de incorporar conhecimento prioritário. A vantagem de tal técnica baseia-se em um princípio relativamente simples, onde ao descorrelacionar os dados, elimina-se parte da informação redundante em cada dimensão. O objetivo da ACP é encontrar uma transformação mais representativa e geralmente mais compacta das observações. De acordo com GONÇALO (2004) o método transforma um vetor aleatório $X \in \mathbb{R}^m$ em outro vetor $y \in \mathbb{R}^n$ (para $n \leq m$) projetando X em n direções ortogonais de maior variância – as componentes principais. Estas componentes são individualmente responsáveis pela variância das observações, e neste sentido, as representam mais claramente. Geralmente, grande parte da variância dos dados é explicada por um número reduzido de componentes, sendo possível descartar as restantes sem perda da informação. De fato, é possível demonstrar que o método de ACP é uma técnica ótima de redução de dimensão linear, relativa ao erro médio quadrático, sendo vantajosa para a compreensão e visualização dos dados, além de reduzir os cálculos necessários em etapas posteriores do processamento dos dados.

Segundo PEREIRA (2005) a ACP estabelece, com base em uma matriz de semelhança (correlações, variâncias-covariâncias ou até de similaridades), um conjunto

de eixos (componentes ou fatores) ortogonais. Estes eixos representam a totalidade da variância dos dados, cada um contribuindo com uma determinada fração. Cada componente corresponde a um autovetor dessa matriz. Assim, com base em uma matriz com m variáveis, serão calculados m autovetores (eixos fatoriais) de comprimento $\lambda_1, \lambda_2, \dots, \lambda_m$ decrescente em razão da sua contribuição à variância total de dados. Esses comprimentos correspondem aos m autovalores (raízes latentes) da matriz. O resultado disso é um sistema reduzido de coordenadas. Desta maneira o primeiro procedimento para redução da alta dimensão dos dados foi aplicar o método de análise de “comunidades” (*cumunality*), introduzida por THURSTONE (1947). Esta técnica procura a parte da variância explicada apenas pelos fatores comuns, excluindo, assim, a parte da variância ligada aos fatores específicos, próprios e exclusivos de cada variável.

Outra abordagem considerada foi o de agrupamento através de análise de *Clusters*. Neste problema de agrupamento, o conjunto de n objetos $X = \{X_1, X_2, \dots, X_n\}$ é separado em grupos de objetos similares. Cada $X_i \in \mathfrak{R}_p$ é um vetor atributo consistindo de p medidas reais que descrevem o objeto. Os objetos são agrupados em grupos não superpostos $C = \{C_1, C_2, \dots, C_k\}$ (C é um grupo), onde k é o número de grupos $C_1 \cup C_2 \cup \dots \cup C_k = X$, $C_i \neq \varnothing$ e $C_i \cap C_j = \varnothing$ para $i \neq j$. Os objetos (casos) separados em cada grupo devem ser mais similares entre si do que entre os objetos de qualquer outro grupo, de maneira que o valor de k pode ser desconhecido. Caso contrário, se k é conhecido, trata-se de um problema de k grupos (*clusters*). No caso tratado neste trabalho, o número de *clusters* é igual ao número de fases do ciclo celular bacteriano portanto, conhecido e em número de 3. A separação dos dados ocorre através de cálculos de similaridades sem que nenhuma suposição sobre a possível estrutura existente seja feita. A similaridade entre objetos (dados) é uma medida que compara quão próximos (ou parecidos) são os objetos em questão. Portanto, uma pequena distância entre objetos deverá indicar uma alta similaridade. Desta maneira, pode-se usar as medidas de similaridade também de forma inversa, como medida de dissimilaridade (diferença). De acordo com EVERITT (1993) e RICHARD (1992) várias medidas de distância podem ser empregadas no problema de agrupamento de dados sendo a mais comum e utilizada neste trabalho a distância euclidiana dada pela equação.

$$d = (X_i, X_j) = \sqrt{(X_i - X_j)'(X_i - X_j)} = \left[\sum_{l=1}^p (x_{il} - x_{jl})^2 \right]^{1/2}$$

(equação 18)

Existe um grande número de algoritmos de agrupamento disponíveis na literatura e a escolha depende do tipo de dados, da aplicação e do objetivo que se quer alcançar. Em geral, os métodos podem ser classificados em métodos hierárquicos, de partição, densidade, baseados em “grid” e em modelos. Pode-se construir uma estrutura em forma de árvore chamada Dendrogama. (HAN & KAMBER, 2001). Neste estudo será usado apenas o método de partição geralmente conhecido como *K-means*.

O algoritmo *K-means* considera uma base de dados de n objetos, e constrói k *clusters* de dados, cada partição representa um grupo e $k \leq n$. Ou seja, o método classifica os dados em k grupos que satisfazem o seguinte critério: todo grupo deve conter no mínimo um objeto, e cada objeto deve pertencer a exatamente um grupo. Isto resulta em alta similaridade dos elementos dentro do grupo e uma baixa similaridade entre os grupos. A similaridade do *cluster* é medida em relação ao valor médio dos objetos em um *cluster*, que pode ser visto como centro de gravidade do *cluster*. O algoritmo *K-means* funciona da seguinte forma: primeiramente, seleciona k objetos de forma randômica, cada um dos quais representando inicialmente uma média ou o centro do *cluster*. Os objetos restantes são designados para um *cluster* que tenha a maior similaridade, baseado numa medida de distância entre o objeto em questão e a média do *cluster*. O algoritmo computa a nova média do *cluster*. Esse processo prossegue até que um critério de parada seja satisfeito. Tipicamente, um critério de erro quadrático é usado, como definido

$$E = \sum_{i=1}^K \sum_{p \in C_i} |p - m_i|^2 \quad (\text{equação 19})$$

onde E é a soma do erro quadrado de todos os objetos na base de dados, p é o ponto no espaço representando um dado objeto, e m_i é a média do *cluster* C_i (ambos p e m_i são multidimensionais). Esse critério tenta fazer os k clusters resultantes o mais compacto e separado possível. Desta forma o algoritmo *k-means* funciona bem quando os *clusters* na base de dados são números compactos e bem separados uns dos outros. O método é relativamente escalável e eficiente no processamento de grandes conjuntos de dados por causa da complexidade computacional do algoritmo ser $O(nkt)$, onde n é o número total de objetos, k é o número de *clusters*, e t é o número de interações. Normalmente, $k \ll n$ e $t \ll n$. O método termina frequentemente em um ótimo local.

O método *k-means*, entretanto, pode ser aplicado quando a média do *cluster* é definida. A necessidade para o usuário especificar o valor de k , o número de *clusters*,

isto pode ser visto como uma desvantagem, porém não é o caso deste trabalho, pois o valor de k é referente ao número de fases do ciclo celular bacteriano. O método *K-means* não é adequado para descobrir *clusters* não convexos, ou *clusters* de forma complexas. Entretanto, é sensível à ruídos e *outliers*.

3.6.2 Redes Neurais

As redes neurais artificiais segundo HAYKIN (2001) são modelos computacionais não lineares inspirados na estrutura e operação do cérebro humano, que procuram reproduzir características tais como: *aprendizado*, *associação*, *generalização* e *abstração*. Estes modelos são compostos por várias unidades de processamento interconectadas (neurônios) que podem adquirir conhecimento através de exemplos. A estas conexões (sinapses) são atribuídos pesos, cujo objetivo é determinar o grau de influência de cada conexão no neurônio artificial (Figura 3.7).

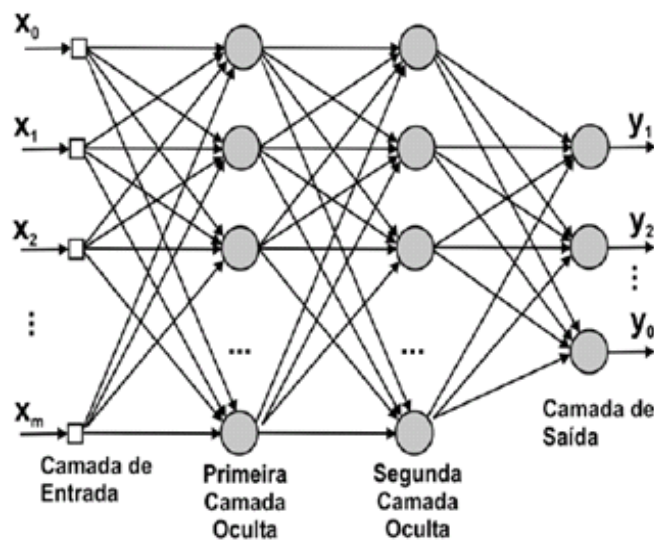


Figura 3.7 – Modelo de rede neural artificial do tipo *feedforward* apresentando uma camada de entrada, duas camadas escondidas e uma camada de saída com três neurônios

Quando se observa o modelo do neurônio artificial (Figura 3.8) identificam-se três elementos básicos:

- As sinapses ou conexões, caracterizadas por uma ponderação ou peso (w_k) que são representadas por um conjunto de retas que convergem como entradas do referido neurônio;

- Um somador (equivalente ao corpo celular) que soma as entradas (X_1, X_2, \dots, X_n), ponderadas pelos seus respectivos pesos sinápticos. As operações aqui constituem um combinador linear.
- A saída (Y_k) representada aqui por uma função de ativação (derivável) para restringir o sinal dentro de uma certa amplitude.

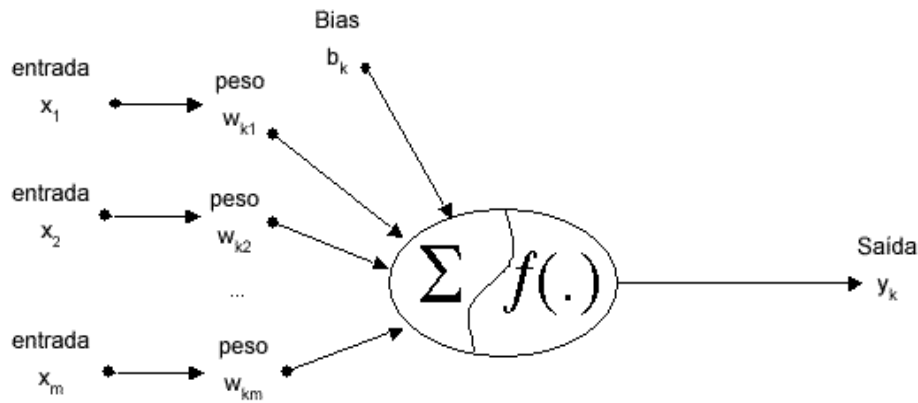


Figura 3.8 – Esquema de um neurônio artificial mostrando suas sinapses com respectivos peso, o somador com a função de ativação e saída.

Nota-se ainda, a existência de uma sinapse chamada “*bias*” (polarização) representada por b_k . O *bias* tem o efeito de aumentar ou diminuir a entrada líquida da função de ativação dependendo se ele é positivo ou negativo.

O modelo do neurônio artificial funciona da seguinte forma. As variáveis de entrada (X_1, X_2, \dots, X_n) que chegam pelas sinapses são multiplicadas pelos seus respectivos pesos ($W_{k1}, W_{k2}, \dots, W_{kn}$) e então são somados no corpo do neurônio. A este resultado é aplicada uma função de transferência chamada de função de ativação do neurônio, resultando em um sinal de saída Y_k . Desta maneira, pode-se descrever matematicamente o neurônio U_k segundo as seguintes equações:

$$U_k = \sum W_{kj} X_j \quad (\text{equação 20})$$

$$Y_k = \varphi(U_k + b_k) \quad (\text{equação 21})$$

onde: X_1, X_2, \dots, X_n são as entradas;

W_1, W_2, \dots, W_n são os pesos sinápticos do neurônio k ;

U_k é a saída do combinador linear;

B_k é o *bias*;

$\varphi(\bullet)$ é a função de ativação;

Y_k é o sinal de saída do neurônio

As funções de ativação mais utilizadas são a sigmoide logística (não-simétrica) e a tangente hiperbólica (assimétrica) cujos valores normalizados de saída do neurônio são escritos intervalo fechado $[0,1]$ e $[-1,1]$ respectivamente. A forma mais utilizada de não linearidade é uma sigmoide definida pela função logística.

$$y_1 = \frac{1}{1 + \exp(-v_1)} \quad (\text{equação 22})$$

onde v_j é o campo local induzido (a soma ponderada de todas as entradas sinápticas acrescidas do bias) do neurônio j , y_j é a saída do neurônio. Porém, de acordo com HAYKIN (2001) o perceptron de múltiplas camadas aprende mais rápido quando a função de ativação incorporada aos neurônios da rede for assimétrica do que quando ela é não-simétrica. Diz-se então que a função de ativação é *anti-simétrica* se:

$$\varphi(-v) = \varphi(v) \quad (\text{equação 23})$$

como representado na Figura 3.9a. Esta condição não é satisfeita pela função logística padrão, representada na Figura 3.9b.

Um exemplo de função de ativação anti-assimétrica é uma não linearidade sigmoide na forma de uma tangente hiperbólica, definida por

$$\varphi(v) = a \tan g(bv) \quad (\text{equação 24})$$

onde a e b são constantes. Os valores adequados para a e b são apresentados na Figura 3.9.

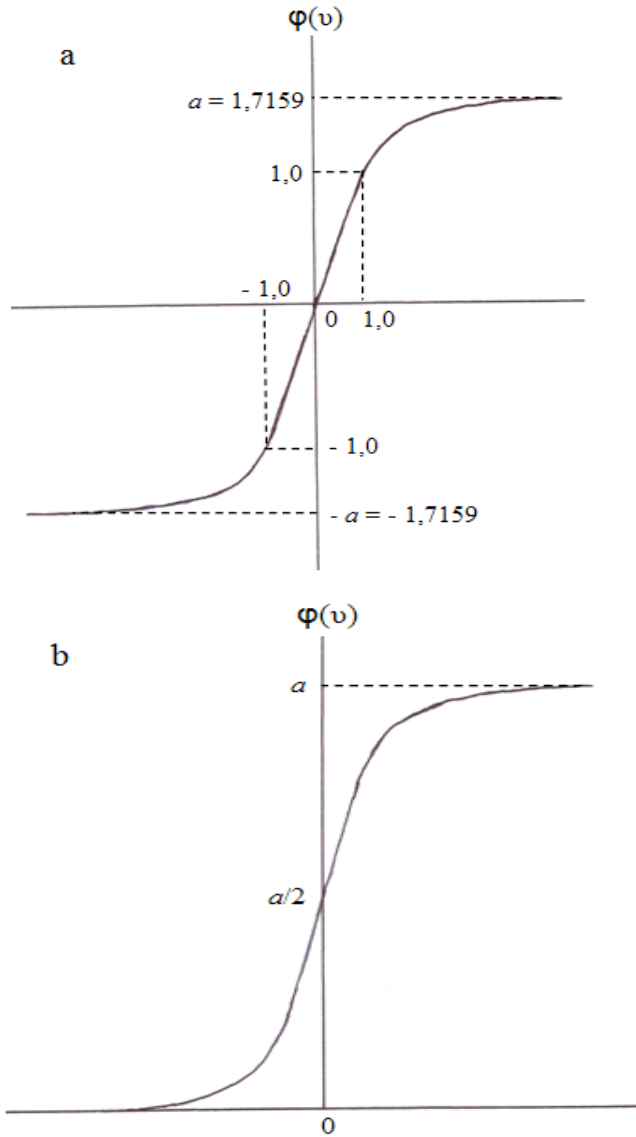
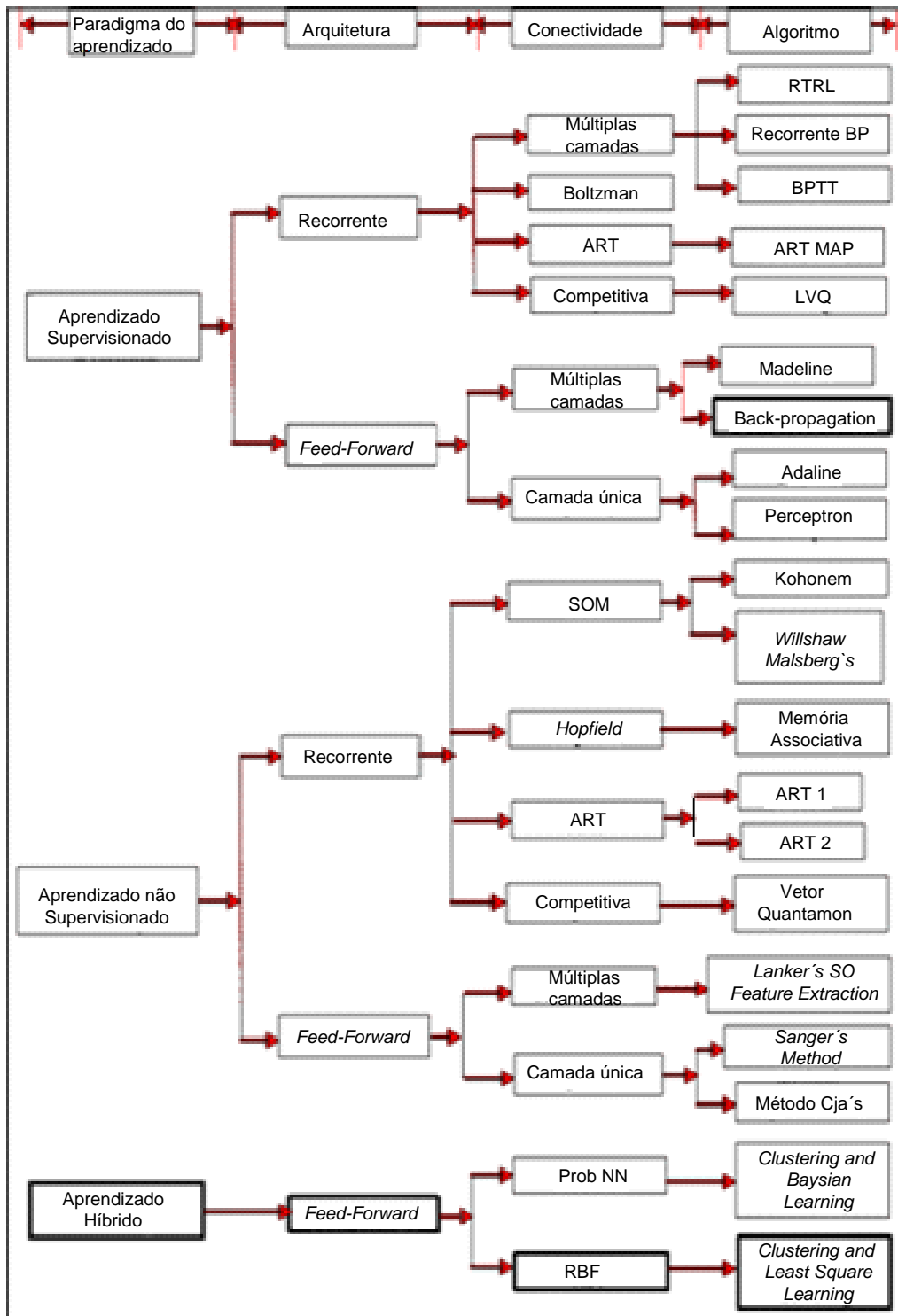


Figura 3.9 – (a) Função de ativação assimétrica, (b) Função de ativação não assimétrica. φ é a função de ativação do neurônio enquanto v é chamado de campo local induzido ou potencial de ativação do neurônio.

Existem muitos tipos de modelos de redes neurais encontrados na literatura. Neste trabalho será adotada a classificação proposta por SUNDARJAN (1998) que se baseia no tipo de treinamento, na arquitetura, na conectividade e nos algoritmos de aprendizado, conforme apresentada na Tabela 3.1.

Tabela 3.1 – Esquema de classificação para redes neurais segundo SUNDARJAN, 1998



Os métodos utilizados para se definir e ajustar os pesos de uma rede são chamados de algoritmos de treinamento: Existem três tipos de treinamento ou paradigmas de aprendizado:

- Supervisionado;
- Não supervisionado;
- Híbrido ou treinamento por reforço.

No treinamento supervisionado, um agente externo (supervisor) indica à rede a resposta desejada para o padrão de entrada. O ajuste dos pesos se dá de acordo com a diferença entre o valor desejado e a saída da rede. Quando o erro atinge um valor satisfatório diz-se que a rede adquiriu conhecimento e considera-se como treinada.

No treinamento não supervisionado, não se dispõe de uma saída alvo, a rede desenvolve a capacidade de representar automaticamente os dados de entrada através de novas classes, interpretando estatisticamente a regularidade dos dados de entrada.

O treinamento híbrido ou treinamento por reforço pode ser considerado uma variação do treinamento supervisionado. Pode-se dizer que existe um “crítico” que verifica se a resposta da rede é satisfatória, em caso afirmativo as sinapses são reforçadas, casos contrários devem ter menor peso.

A definição da arquitetura de uma rede neural artificial é um parâmetro importante na sua concepção, uma vez que ela restringe o tipo de problema que pode ser trabalhado pela rede, resultando no seu sucesso ou fracasso. Uma rede neural pode ser classificada quanto a sua arquitetura como:

- Redes alimentadas adiante (feedforward). Nestas redes as saídas dos neurônios de uma determinada camada somente se conectam com os neurônios da camada subsequente
- Redes recorrentes (feedback). Nestas redes, pelo menos uma das conexões sinápticas está ligada a algum neurônio de uma camada anterior ou a ele mesmo.

A maneira pela qual os neurônios estão estruturados define a topologia neural e está intimamente ligada ao algoritmo de aprendizado.

Devido as entradas das redes neurais artificiais serem conhecidas, e as funções de ativação escolhidas pelo usuário, a saída total pode ser calculada como uma função dos pesos. É através do processo de ajuste desses pesos que a rede aprende e adquire conhecimento. Em termos do modelo artificial a conectividade é expressa pelos padrões de ligação entre os neurônios e aos pesos atribuídos a essas ligações.

As possibilidades de ligações entre os neurônios são imensas. A maneira pela qual ocorrem as conexões, o número e a disposição dos neurônios da rede chamam-se topologia.

Uma rede neural pode ser classificada por sua conectividade como:

- Completamente conectada, quando todos os neurônios de uma camada estão conectados a todos os neurônios da camada posterior.
- Parcialmente conectada, caso falte alguma conexão entre os neurônios da camada subsequente.

As conexões podem ser:

- Laterais, entre os neurônios da mesma camada;
- Intercamadas, entre neurônios de diferentes camadas;
- Autoexcitatórias, partem e atingem o mesmo neurônio.

Em aplicações cujos problemas são de classificação, o objetivo de uma rede neural artificial é designar cada caso a um número de classes, mas geralmente, estimar a probabilidade de pertinência de um determinado caso para uma determinada classe. Normalmente são usadas duas estratégias. A primeira trata de variáveis de saída consideradas como *dois-estados* (*two-state*), a outra técnica considera a variável de saída como *um-de-N* (*one-of-N*). Na primeira representação, um único nó (neurônio) corresponde a variável, e um valor de 0 é interpretado como um estado, e um valor de 1 como um outro estado. Na segunda, uma unidade de saída é alocada para cada estado.

O nível de confiança da rede decide se aceita ou rejeita o nível de ativação (*thresholds*) dos neurônios, auxiliando na interpretação das saídas da rede. Os *thresholds* podem ser ajustados, a fim de tornar a rede mais ou menos “precisa” (*fuzzy*) em relação a designação de uma classificação. A interpretação difere um pouco entre as duas representações. Na abordagem *Two-state*, se a unidade de saída está acima do *threshold* designado a classe 1 é escolhida, caso contrário, a classe 0 será escolhida. Na abordagem *One-of-N*, a classe é selecionada se a unidade de saída correspondente está

acima do *threshold* aceito e todas as outras unidades de saída estão abaixo deste *threshold*.

Este trabalho utilizou um modelo de redes neurais denominado Perceptron Multicamadas (*Multi Layer Perceptron* - MLP) treinado com o algoritmo BFGS. O Algoritmo BFGS chamado método de otimização de Broyden-Fletcher-Goldfarb-Shanno (AVRIEL, 2003) é um método de otimização não-linear, obtido a partir de uma variação do método de Newton. O método de Newton assume que a função pode ser localmente aproximada como uma função quadrática e busca dessa forma, o ponto de estacionariedade ou de derivada nula.

Nos métodos de *Quase-Newton*, a matriz Hessiana de derivadas de segunda ordem da função a ser minimizada não necessita ser calculada. Alternativamente, a matriz Hessiana também é estimada partindo-se de uma matriz inicial, normalmente a identidade, e atualizado iterativamente os seus elementos a partir dos vetores gradiente locais. O algoritmo iterativo do método BFGS é apresentado a seguir.

A partir de uma suposição de um X_0 inicial e uma matriz Hessiana B_0 aproximada as seguintes etapas são repetidas até x convergir para a solução.

1. Obter a direção P_k pela solução de:

$$B_k P_k = -\nabla f(X_k) \quad (\text{equação 25})$$

2. Executar uma busca de linha (*line search*) para achar um *stepsize* aceitável α_k na direção encontrada na primeira etapa, então atualizar,

$$X_{k+1} = X_k + \alpha_k P_k \quad (\text{equação 26})$$

3. Estabelecer

$$S_k = \alpha_k P_k \quad (\text{equação 27})$$

4. $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ (equação 28)

$$5. B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T S_k} - \frac{B_k S_k (B_k S_k)^T}{S_k^T B_k S_k} \quad (\text{equação 29})$$

$f(x)$ denota a função objetivo a ser minimizada. A convergência pode ser averiguada observando-se a norma do gradiente, $|\nabla f(x_k)|$. Praticamente, B_0 pode ser inicializado com $B_0 = I$, tanto que a primeira etapa será equivalente ao gradiente descendente, mas,

as etapas posteriores são mais e mais refinadas por B_k , a aproximação da matriz Hessiana. A primeira etapa do algoritmo é efetuada usando-se uma aproximação da matriz inversa B_k , que é normalmente obtida eficientemente pela aplicação da fórmula de Sherman–Morrison para a quinta linha do algoritmo dando:

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(s_k^T y_k + y_k^T B_k^{-1} y_k)(s_k s_k^T)}{(s_k^T y_k)^2} - \frac{B_k^{-1} y_k s_k^T + s_k y_k^T B_k^{-1}}{s_k^T y_k} \quad (\text{equação 30})$$

Os intervalos de confiança para a solução podem ser obtidos a partir da matriz Hessiana inversa final.

O processo de definição da arquitetura neural e sua conectividade é inteiramente empírico e específica sua aplicação, além de determinar seu sucesso ou fracasso. Portanto, é necessária a automação desta tarefa. Técnicas de otimização tradicionais podem ser usadas para definir o desenho do sistema neural, mas segundo JAYALAKSHIMI, (2000) a maior desvantagem destas técnicas é que terminam em um ótimo local. Os algoritmos genéticos são ditos capazes de achar uma otimização global, portanto serão usados neste trabalho para a tarefa da escolha da melhor estratégia de redes neurais.

3.6.3 Algoritmos genéticos

Algoritmos genéticos – AG’s (HOLLAND, 1975) são modelos estocásticos e probabilísticos, inspirados na teoria da evolução natural das espécies e aplicados a problemas complexos de otimização. Um problema de otimização caracteriza-se fundamentalmente por encontrar uma solução entre um número muito grande de possíveis soluções (espaço de busca). Desta maneira, aplicam-se operadores genéticos (seleção natural, cruzamento e mutação) sobre uma população inicial e suas próximas gerações. Cada indivíduo (topologia de rede) da população é codificado em um cromossomo, que representa uma possível solução do problema. A adaptação de um indivíduo é então avaliada pelo valor da função objetivo (seleção natural) e os indivíduos mais adaptados são permitidos a se reproduzirem pela troca de informação genética (cruzamento) com outro indivíduo também adaptado, produzindo indivíduos ainda melhores. Após a criação da população inicial (em geral, aleatória), tem início um processo iterativo de refinamento ou evolução das soluções iniciais.

O AG cria novas soluções através da combinação e refinamento das informações dos cromossomos usando operadores de seleção, *crossover* e mutação. Essas operações produzem novas soluções que formam uma nova população. Cada nova população é chamada de geração. As mutações são frequentemente permitidas pela alteração de alguns genes dos cromossomos. Os melhores indivíduos podem substituir toda geração anterior ou substituir apenas os indivíduos menos adaptados. Geralmente, os AGs proporcionam excelentes resultados na procura de ótimos globais.

Uma das vantagens de um algoritmo genético é a simplificação que eles permitem na formulação e solução de problemas de otimização. Um AG simples normalmente trabalha com descrições da entrada formadas por cadeias de bits de tamanho fixo, porém outros trabalham com cadeias de bits de tamanho variável. Os AGs possuem um paralelismo implícito decorrente da avaliação independente de cada uma dessas cadeias de bits, ou seja, pode-se avaliar a viabilidade de um conjunto de parâmetros para a solução do problema de otimização em questão. Os AG's são numericamente robustos, ou seja, não são sensíveis a erros de arredondamento no que se refere aos seus resultados finais.

Existem três tipos de representação possíveis para os cromossomos: binária, inteira ou real. De acordo com a classe de problema que se deseja resolver pode-se usar qualquer um dos três tipos.

Uma implementação de um algoritmo genético começa com uma população aleatória de cromossomos. Essas estruturas são avaliadas e associadas a uma probabilidade de reprodução de tal forma que as maiores probabilidades são associadas aos cromossomos que representam uma melhor solução para o problema de otimização do que àqueles que representam uma solução pior. A *aptidão* da solução é tipicamente definida com relação à população corrente.

A função objetivo de um problema de otimização é construída a partir dos parâmetros envolvidos no problema. Ela fornece uma medida da proximidade da solução em relação a um conjunto de parâmetros. O objetivo é encontrar o ponto ótimo. A função objetivo permite o cálculo da *aptidão bruta* de cada indivíduo, que fornecerá o valor a ser usado para o cálculo de sua probabilidade de ser selecionado para reprodução.

Como técnica de busca e otimização os AGs apresentam:

- Um espaço de busca, onde estão as possíveis soluções para o problema,

- Uma função objetivo, que associa a cada cromossomo uma aptidão (nota).

O espaço de busca é o conjunto de todas as configurações que um cromossomo pode assumir. Os cromossomos por sua vez, são estruturas de dados, geralmente vetores ou cadeias de bits, (possível solução para o problema de otimização) que representam os parâmetros da função objetivo.

Cada elemento do vetor cromossomo é chamado de gene, estes determinam as características do ser; sendo a codificação binária (0 e 1) a forma mais comum de representá-los.

A função objetivo, também chamada de função de adequabilidade (*fitness function*) quantifica a adaptabilidade de cada cromossomo como uma solução, e é usada como base para selecioná-los para reprodução.

Portanto, o conjunto de cromossomos forma o que se chama de população e as interações do algoritmo genético são chamadas gerações, onde, durante este processo, os operadores genéticos dos AGs agem no material genético. De acordo com Cole, (1998) existem quatro operadores genéticos:

Inicialização, seleção, recombinação e mutação.

- O operador de inicialização é usado para gerar a população inicial para o AG. Esta população deve conter cromossomos que estejam espalhados no espaço de soluções (busca), suprindo o AG de diversificado material genético. A maneira mais fácil de conseguir isto é uma seleção randômica dos cromossomos,
- O operador de seleção escolhe os indivíduos para reprodução baseado em suas funções de adequabilidade. Existem diversos métodos para a seleção dentre os quais os mais conhecidos são a Roda da Roleta que utiliza os operadores de recombinação (*crossover*) e mutação, ou, seleção por Elitismo.
- O operador de recombinação combina o material genético de um ou mais cromossomos pais produzindo uma nova população de cromossomos filhos.
- O operador de mutação introduz novo material genético na população. Em uma codificação binária, por exemplo, simplesmente pode-se mudar um 0 em 1 ou, um 1 em 0 em qualquer ponto do cromossomo.

Na sua forma mais simples (Figura 3.11), pode-se descrever o Algoritmo Genético como:

- 1 – Codifica-se o problema para gerar uma população inicial,
- 2 – Enquanto certo critério de parada não for satisfeito, faça-se:
 - a – Calcular a aptidão de cada indivíduo na corrente população
 - b – Selecione os indivíduos de alta aptidão e fazer cópias dos selecionados,
 - c – Aplicar os operadores de recombinação e mutação criando nova geração.
- 3 – Fim.

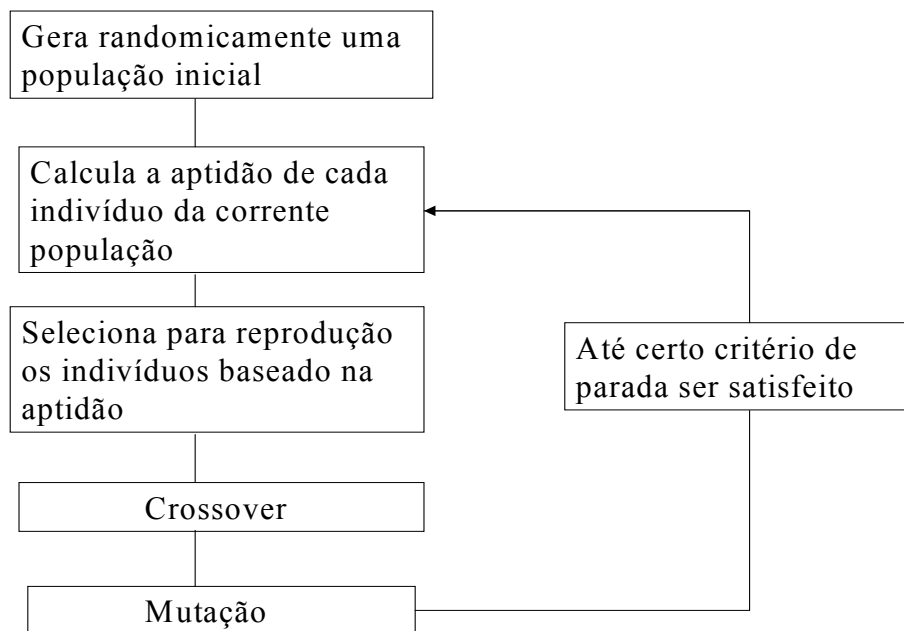


Figura 3.11 – Perfil de um simples Algoritmo Genético

Desta maneira os dados citométricos obtidos e tratados por análises estatísticas, clusterização e classificação utilizando redes neurais foram utilizados para gerar o modelo de classificação de padrões celulares bacterianos, que está apresentado a seguir.

CAPÍTULO IV

Análise dos Resultados e Discussão

Neste capítulo apresentam-se os resultados obtidos no monitoramento da cultura da bactéria *E coli* DH10b pela técnica de citometria de fluxo na tentativa de visualizar os diferentes grupos celulares distribuídos ao longo do ciclo celular, os resultados dos métodos estatísticos empregados na fase de pré-processamento dos dados e o desempenho dos modelos de redes neurais artificiais na tarefa de classificação dos padrões celulares .

A linhagem celular DH10b cultivada sob as condições descritas no item 3.2 do Capítulo III apresentou uma curva de crescimento como demonstra a Figura 4.1, de acordo com as medidas citométricas e de densidade ótica.

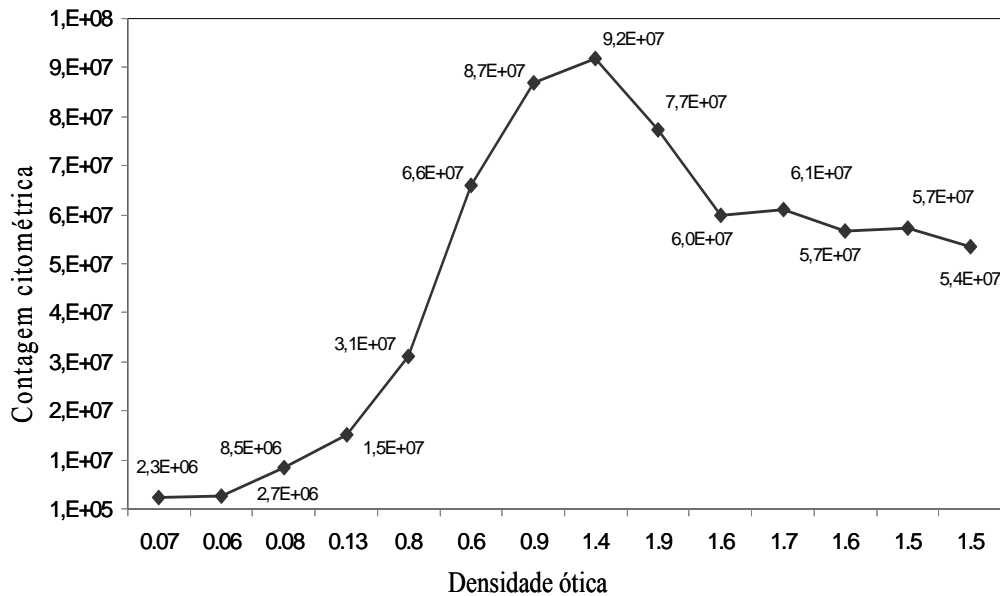


Figura 4.1 – Curva de crescimento de *E. coli* a 37 °C

O primeiro ponto da curva foi realizado apenas 30 minutos após o inóculo, com uma concentração inicial de 2344 células/mL (2,34E+03) e uma densidade ótica de 0.07. Diferentemente da curva padrão apresentada na Figura 2.5, verificou-se que neste experimento a DH10b não apresentou fase estacionária propriamente dita. Foi observado após 4 horas de crescimento (T8) uma concentração máxima de 9,20E+07 e

logo após uma fase de pequeno declínio ($6,00E+07$) devido ao ajustamento da cultura as novas condições do meio, já que se trata de uma cultura do tipo *batch*. Contudo, nos tempos seguintes até o término das contagens observou-se uma tendência a estabilidade.

A cultura apresentou um tempo de geração que variou de 3,17 a 11,44 e uma taxa de crescimento que variou de 0,31 a 0,08. Na realidade, a medida em que a cultura avançou no tempo, ambos os parâmetros, os tempos de geração e taxas de crescimento, mostraram-se inversamente proporcionais. Isto se deve a depleção dos nutrientes, o aumento de competição entre os indivíduos e o efeito da densidade populacional.

Segundo HOLT *et al.*(1996) a *E.coli* apresenta dimensões que variam entre 1,6 e 7,6 μm de comprimento e 0,8 e 1,6 μm de largura. Para acessar a população bacteriana pela técnica de citometria de fluxo foram utilizadas microesferas com os tamanhos de 0,95 μm , 1,6 μm e 6,0 μm de diâmetro. As microesferas de 0,95 e 6,0 μm apresentavam fluorescência compatível com a emissão da molécula do fluorocromo SYBR Green I na faixa de 587 a 653 nanômetros, correspondente ao sensor de FLO do citômetro CytoSub. A Figura 4.2 apresenta o citograma destas microesferas de acordo com seus valores máximos de SWS (eixo X) e FWS (eixo Y) e identificadas por retângulos (*gates*) separadamente. O aparelho foi pré-programado para uma aquisição de dados com o sensor SWS (*trigger channel*) Pode-se observar ainda muitos sinais de diversas e pequenas partículas e sinais considerados como ruído.

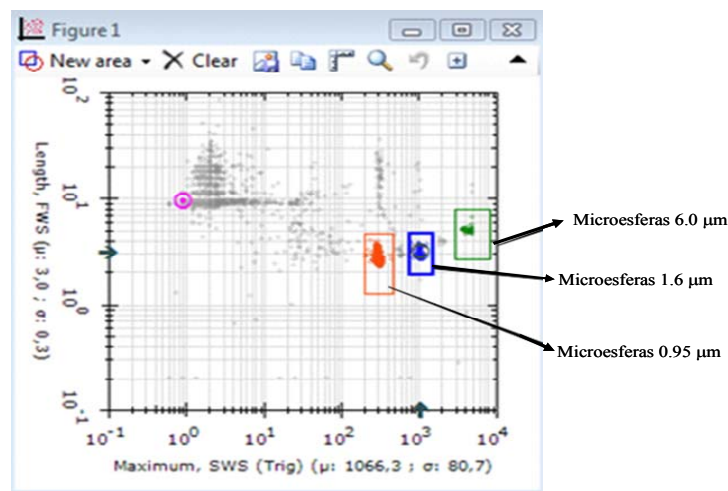


Figura 4.2 – Distribuição de microesferas de 6, 1.6 e 0,95 micrômetros segundo os valores máximos de SWS e *Length* FWS.

Estas mesmas microesferas são apresentadas na Figura 4.3 sem os sinais de ruído e das partículas sem interesse direto para o estudo. Esta possibilidade é um recurso computacional oferecido pelo programa de análise CytoClus 3.

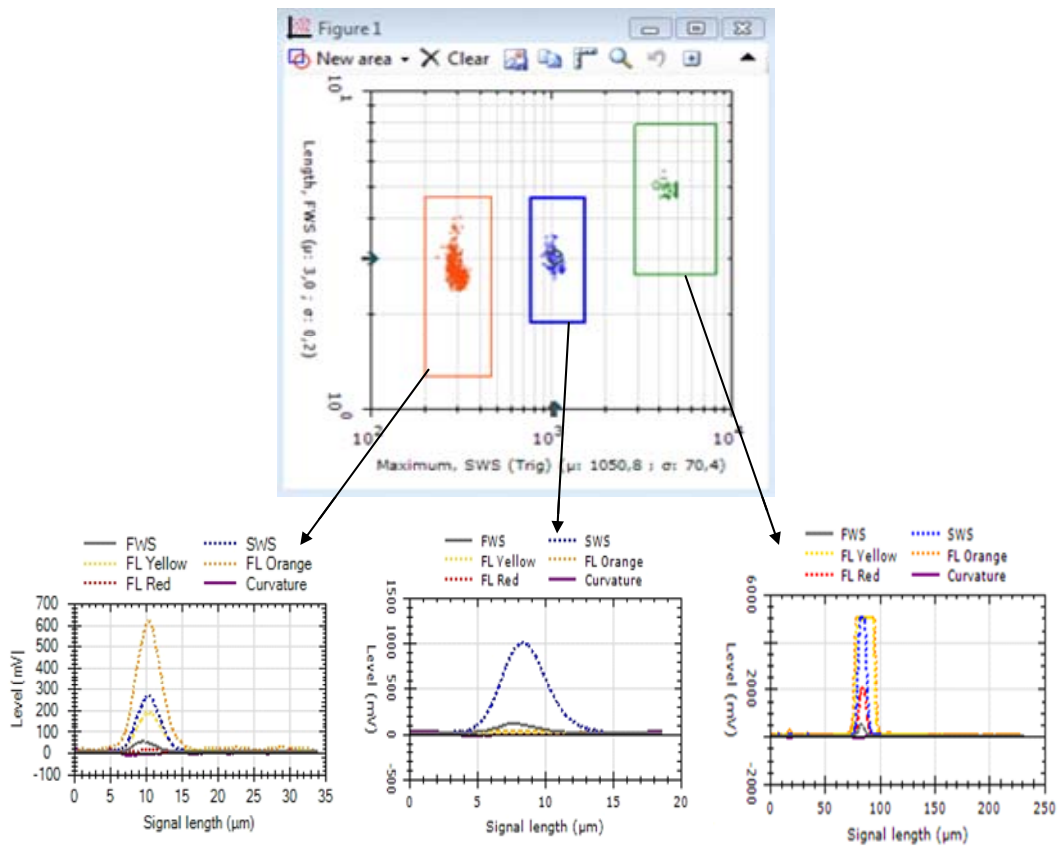


Figura 4.3 - Distribuição de microesferas de 6, 1.6 e 0,95 micrômetros pelos parâmetros SWS e FWS e suas respectivas assinaturas óticas.

O conjunto do perfil das microesferas coletado por cada sensor, assinatura ótica (*pulse shape*), pode ser observado. A esquerda a assinatura ótica característica das microesferas de 0,95 μm apresenta um pulso de fluorescência laranja bem definido com intensidade pouco acima de 600 mV. Este padrão é o ideal para os procedimentos de calibração do canal de fluorescência FLO e dos dados provenientes deste. Diferentemente, do lado direito é apresentada uma assinatura ótica das microesferas de 6 μm . Esta assinatura ótica encontra-se com o sinal saturado (5000 mV) para a fluorescência laranja portanto, não pode ser utilizada para calibração deste sensor porém, ao mesmo tempo, apresenta um pulso bem definido para a fluorescência

vermelha que no caso deste trabalho, sem utilidade. A assinatura óptica referente as microesferas de 1,6 μm apresenta um sinal de SWS sem qualquer pulso de fluorescência, neste caso só pode ser utilizada para calibrações de tamanho de partículas.

O sinal de FWS é o mais usado para obtenção de informações relativas ao comprimento de partículas. Porém, para partículas com dimensões muito pequenas, como as bactérias, é comum encontrar na literatura científica o uso do sinal de SWS para esta tarefa.

Na Figura 4.4a observa-se que todas as microesferas encontram-se agora agrupadas pela razão length do SWS/Fill factor SWS (eixo X), e pela razão length FWS/TOF SWS (eixo Y) em uma região demarcada pelo retângulo (*gate*) amarelo que corresponde a faixa de variação de tamanho da *E coli*. Este retângulo foi então transferido automaticamente para outro arquivo cujos dados eram provenientes da cultura de bactérias em sua fase exponencial, como demonstrado na Figura 4.4b.

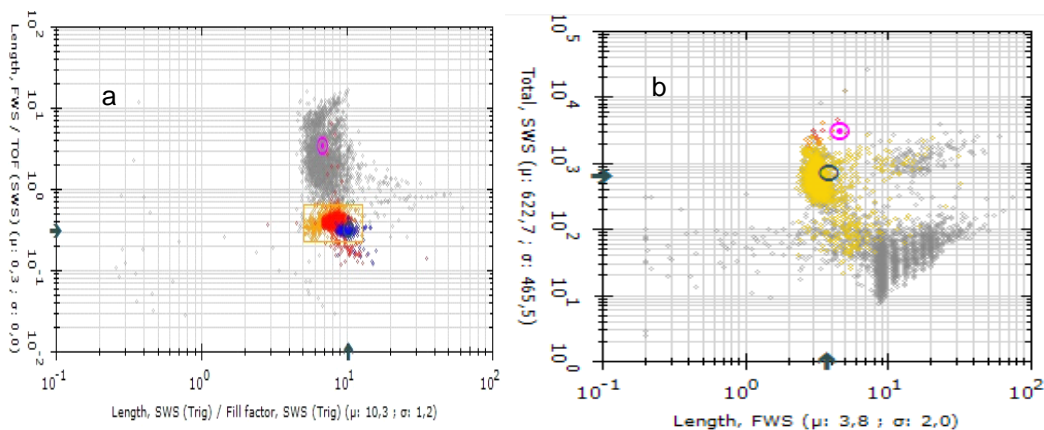


Figura 4.4 – Em a, microesferas de 0,95 e 1,6 micrômetros agrupadas segundo a razão Length FWS/TOF SWS no eixo y, em b o retângulo delimitado em A é transposto para mostrar apenas bactérias.

A região observada na Figura 4.4b apresenta três grupos de células coloridas em amarelo devido a cor escolhida para o retângulo na Figura 4.4a, correspondendo aos três *clusters* de bactérias além dos dados considerados como ruídos. Novamente os recursos disponibilizados pelo programa CytoClus 3 foi utilizado como estratégia de filtro de ruídos.

Desta maneira a Figura 4.5 apresenta os mesmos três grupos de células bacterianas de acordo com os sinais dos parâmetros length SWS (eixo X) e máximo de fluorescência laranja (eixo Y), além de suas respectivas assinaturas ópticas características para cada grupo celular. Cabe ressaltar que estes dados referem-se ao T5 da curva de crescimento apresentada na Figura 4.1, momento em que a cultura apresentou o maior nível de ativação celular, maior taxa de crescimento e menor tempo de geração. Isto pode ser observado pelas assinaturas ópticas de cada um dos *clusters*, onde a assinatura referente ao grupo verde (*cluster 1*), de menor densidade populacional (5,12 %), apresenta um nível de fluorescência laranja de 60 mV (em baixo), enquanto que o grupo azul (*cluster 2*), de maior densidade populacional (74,46 %), o nível de 120 mV (esquerda). O grupo vermelho (*cluster 3*), com densidade populacional intermediária (20,41 %), apresenta dois picos de fluorescência (direita), o primeiro acima de 200 mV e o segundo em 400mV, o que sugere que este grupo encontra-se em estágio terminal do processo de divisão celular.

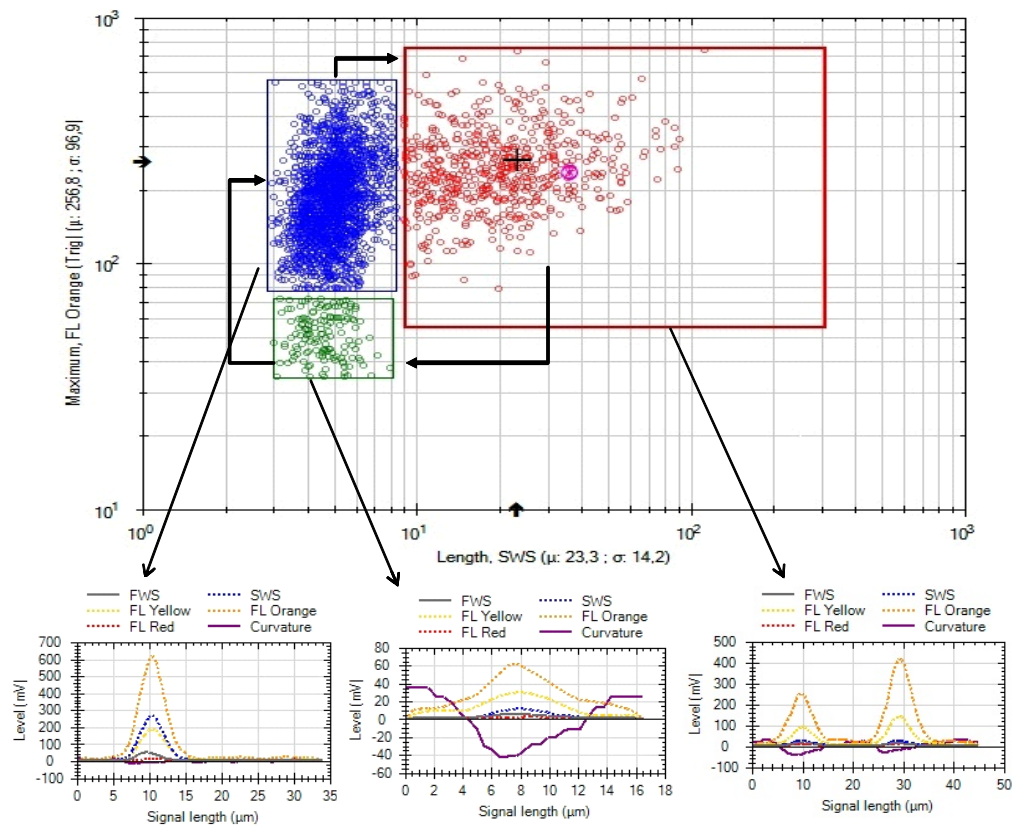


Figura 4.5 – Citograma reduzido (sem ruído) apresentando os dois grupos de *E. coli* com DNA marcados. A direita, células em processo de divisão, a esquerda, células únicas.

Estes resultados sugerem que o grupo verde é composto por células que se encontram na chamada fase B do ciclo celular (SKARSTAD *et al.*, 1983), que estão no tempo compreendido entre a divisão celular e a inicialização de um novo processo de duplicação cromossômica. Neste sentido as células do grupo azul, encontram-se na fase C do ciclo celular caracterizada por indivíduos que já possuem seu material genético duplicado enquanto, o grupo vermelho é composto por células na fase D do ciclo celular, término da duplicação cromossômica e início da duplicação celular. O término da síntese de DNA é o sinal para a deposição do anel de FtsZ, que é mediado pelas proteínas Min, sendo responsável pela constricção da membrana e parede celulares, como mencionado anteriormente no Capítulo II.

A Figura 4.6 apresenta a progressão temporal do conteúdo de DNA na população da linhagem DH10b em diferentes tempos da curva de crescimento.

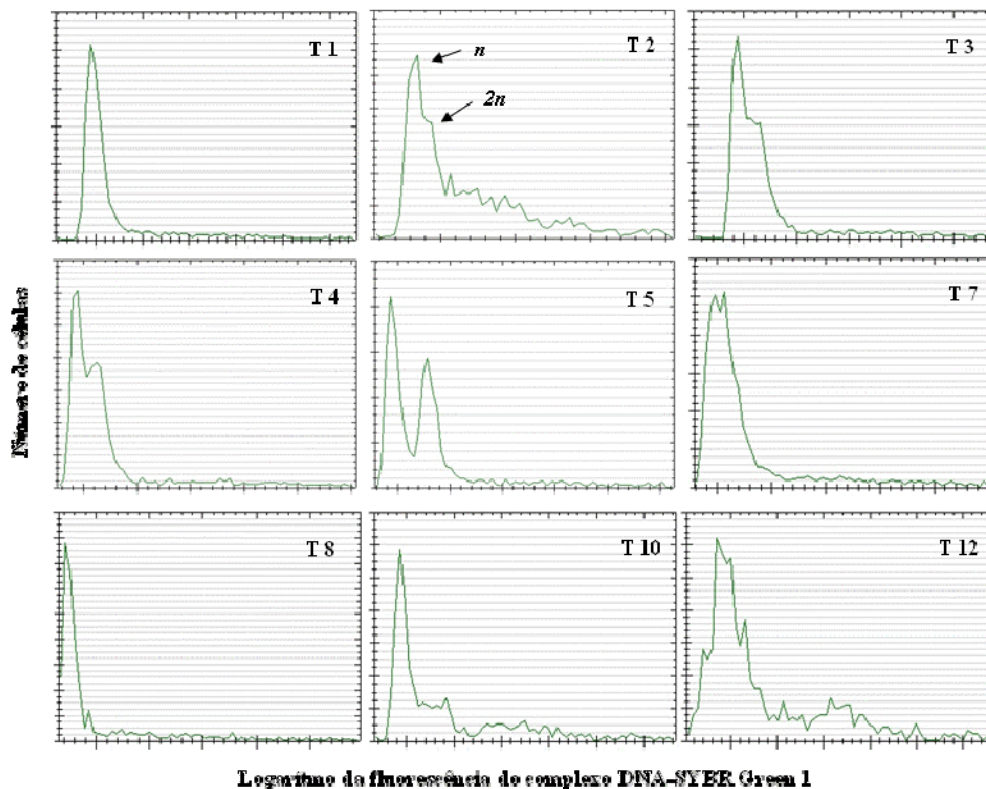


Figura 4.6 – Histogramas da distribuição temporal do número de células com DNA marcado com fluorocromo SYBR Green I ao longo da curva de crescimento.

Observa-se que o tempo T1 apresenta um único pico de fluorescência laranja sugerindo que as células de DH10b constituem uma população bastante homogênea (n). Em T2, o mesmo pico começa a apresentar alguma diferenciação. Neste momento pode-se perceber a formação de uma nova população ($2n$). O mesmo é verificado em T3 porém com maior intensidade. Estes tempos correspondem a fase lag da curva de crescimento. No início da fase exponencial, T4, o grupo de células $2n$ começa a aumentar significativamente tornando-se bem visível e diferenciado em T5 e equivalente ao grupo de células n no tempo T7. Em T8, ponto máximo de crescimento, o grupo $2n$ diminui sensivelmente. A partir deste momento a cultura passa por ajustamento as novas condições ambientais.

Foi possível diferenciar nível de atividade metabólica individual medida de acordo com a citometria de fluxo e o uso do diacetato de carboxifluoresceína (CFDA). Quando as células encontram-se metabolicamente ativas produzem enzimas que clivam o CFDA, dando como resultado a emissão de fluorescência na faixa de 525 nm, correspondente ao sensor de fluorescência laranja do CytoSub. Na Figura 4.7 observa-se a sequência de eventos para células com atividade metabólica e sem atividade metabólica.

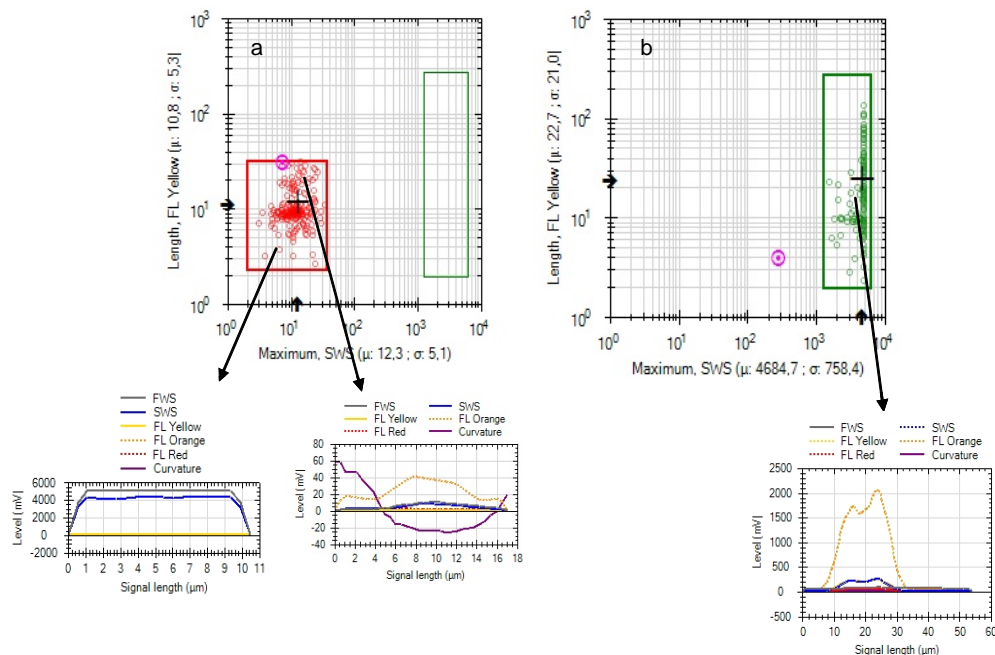


Figura 4.7 – Em a observa-se o controle negativo para a atividade metabólica, quadrante vermelho e as possíveis assinaturas óticas. Em b o quadrante verde mostra as células com atividade e sua respectiva assinatura ótica.

Pode ser verificado na Figura 4.7 a que as células sem atividade metabólica ou com pouca atividade metabólica apresentam a assinatura ótica com baixa ou nenhuma fluorescência na faixa do amarelo e a área retangular relativa as células com atividade encontra-se vazio, pois o tratamento com álcool inibiu a atividade metabólica. Já na Figura 4.7b a região que representa o grupo de células com atividade enzimática encontra-se com uma assinatura ótica com fluorescência apesar de possuírem SWS e FWS altos. Com a determinação das regiões e assinaturas óticas das subpopulações celulares passou-se a analisar as células mantidas em cultura durante o experimento.

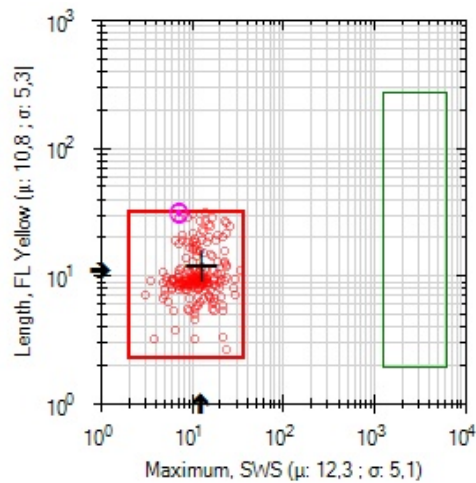


Figura 4.8 - Atividade metabólica da cultura na fase de declínio ou morte celular.

A Figura 4.8 corresponde a fase de declínio da curva de crescimento, onde pode ser observado que a maior parte da população bacteriana encontra-se em um grupo com pouca atividade metabólica (grupo vermelho) e o grupo verde mostra que ainda existem células metabolicamente ativas. Nesta etapa o percentual de células inativas foi de mais de 95 %.

Na Figura 4.9 pode ser observado a distribuição dos grupos bacterianos em relação a atividade metabólica na fase exponencial do crescimento no tempo T7, período de maior taxa de crescimento.

Na fase exponencial do crescimento o percentual de células metabolicamente ativas foi de mais de 98%.

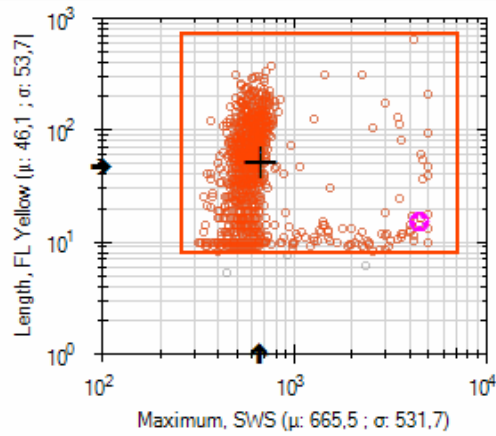


Figura 4.9 – Atividade metabólica da cultura na fase exponencial do crescimento.

Após a fase de monitoramento da cultura, aquisição e análise das assinaturas citométricas e determinação do nível de viabilidade celular passou-se a explorar os dados gerados pela citometria de fluxo. Desta forma foi efetuada uma análise de componentes principais, para demonstrar a distribuição das 47 variáveis obtidas pelo citômetro de fluxo CytoSub de acordo com os eixos fatoriais 1 e 2 (Figura 4.10).

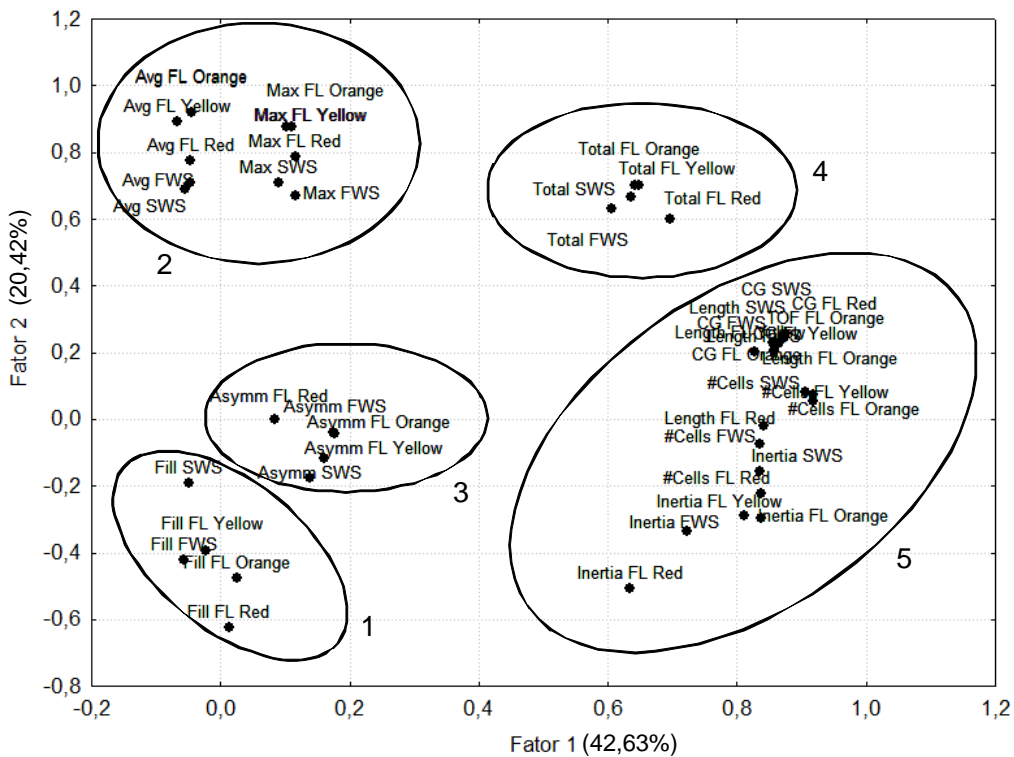


Figura 4.10 – Visualização da distribuição das variáveis pela Análise de Componentes Principais de acordo com os eixos fatoriais 1 e 2.

Pode ser facilmente observado a presença de 5 agrupamentos, resultantes na aplicação do método *varimax normalized*. De acordo com a Tabela 4.1 estas 47 variáveis apresentam apenas 74,37% de variância total explicada em 3 componentes principais. A primeira contribuindo com 42,63%, a segunda com 20,42% e a terceira com 11,32% respectivamente.

Tabela 4.1 – Autovalores das 3 componentes principais.

Valores	Autovalores	Variância Total (%)	Autovalores acumulados	Variância Acumulada (%)
1	20,03662	42,6311	20,03662	42,63110
2	9,5979	20,42107	29,63452	63,05217
3	5,32086	11,32097	34,95538	74,37314

Com o objetivo de melhorar o desempenho dos algoritmos de identificação e classificação das subpopulações bacterianas utilizou-se na fase de pré-tratamento dos dados um estudo de *comunality* a fim de verificar a possibilidade de conseguir uma redução de dimensionalidade da matriz. Desta maneira, foram identificadas as variáveis que mais contribuíram para a solução do problema matemático. Estas variáveis encontram-se na Tabela 4. 2.

Tabela 4.2.- Variáveis selecionadas pela aplicação do método de *comunality*. As variáveis com valores em negrito são as que mais contribuem para cada fator.

Variáveis	Fator 1	Fator 2
TOF FL Orange	0,970172	0,09422
Avg FL Yellow	0,004804	0,984495
Avg FL Orange	0,050371	0,982129
CG FWS	0,975516	0,07986
CG SWS	0,970446	0,084592
CG FL Yellow	0,974164	0,08388
CG FL Orange	0,979342	0,083095
CG FL Red	0,983683	0,088636
#Cells SWS	0,953467	-0,071377
# Cells FL Yellow	0,951198	-0,106187
# Cells FL Orange	0,961549	-0,086293
Var. Expl.	8,45138	2,001833
Perc. Total	0,768307	0,181985

Pode ser observado que houve uma redução significativa no número de variáveis envolvidas no problema (36 variáveis) além da perda da terceira componente principal inicialmente identificada. Na realidade as variáveis que mais contribuem para o Fator 1 são *Time of flight* da fluorescência laranja (TOF FL Orange), que na realidade expressa a fluorescência do complexo Syber Green I/DNA, os valores relativos aos centros de gravidade das variáveis forward scatter (CG FWS), side scatter (CG SWS), fluorescência amarela (CG FL Yellow), fluorescência laranja (CG FL Orange) e fluorescência vermelha (CG Red); os valores relativos ao número de células das variáveis *side scatter* (# Cells SWS), fluorescência amarela (# Cells FL Yellow) e fluorescência laranja (# Cells FL Orange). Os valores médios das fluorescências amarela (Avg Yellow) e laranja (Avg Orange) encontram-se formando o eixo fatorial 2.

Este procedimento de redução de dimensionalidade não promoveu qualquer perda de informação. Percebe-se na Tabela 4.3, um aumento significativo nos valores referentes às variâncias dos eixos fatoriais 1 (76,96%) e 2 (18,06%), e conseqüente aumento da porcentagem de variância total explicada que passou para 95,02%. Na realidade, este procedimento resultou em ganho de informação.

Tabela 4.3 – Autovalores das 2 componentes principais.

Valores	Autovalores	Variância Total (%)	Autovalores acumulados	Variância Acumulada (%)
1	8.465663	76.96057	8.46566	76.96057
2	9,5979	18.06863	10.45321	95.02921

Fica claro a diminuição de 5 para 2 agrupamentos de variáveis. Percebe-se desta maneira que apenas os valores médios de fluorescência e os centros de gravidade de partículas contendo mais de uma célula (# *Cell*) são significativos. Este fato pode ser devido ao sistema de agitação da cultura.

A Figura 4.8 apresenta assim a nova configuração das variáveis selecionadas distribuídas segundo os eixos fatoriais 1 e 2.

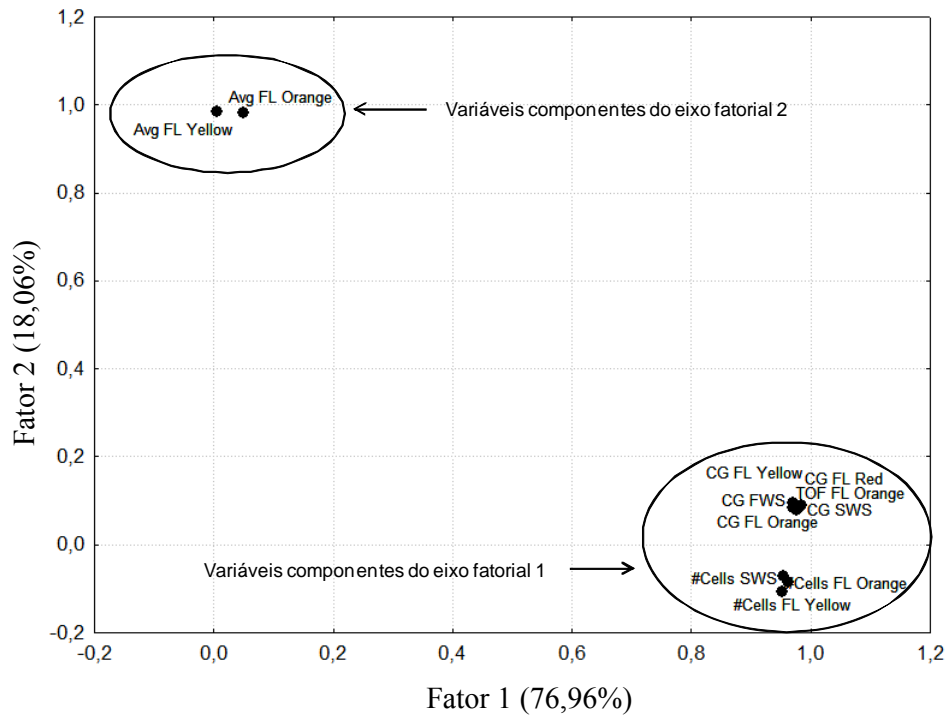


Figura 4.8. - Distribuição das variáveis selecionadas após o estudo de comunidade de acordo com os eixos fatoriais 1 e 2.

A Tabela 4.4 apresenta a correlação linear entre as 11 variáveis selecionadas após o processo de redução da dimensão dos dados.

Tabela 4.4 - Matriz de correlação das variáveis selecionadas
Valores significantes em $p < .05$

Variáveis	TOF FL Orange	Avg FL Yellow	Avg FL Orange	CGFWS	CGSWS	CG FL Yellow	CG FL Orange	CG FL Red	#Cells SWS	#Cells FL Yellow	#Cells FL Orange
TOF FL Orange	1.00										
Avg FL Yellow	0.1	1.00									
Avg FL Orange	0.14	0.94	1.00								
CGFWS	0.94	0.08	0.12	1.00							
CGSWS	0.93	0.08	0.12	0.99	1.00						
CG FL Yellow	0.93	0.08	0.12	0.98	0.99	1.00					
CG FL Orange	0.94	0.08	0.12	0.98	0.99	1	1.00				
CG FL Red	0.96	0.09	0.13	0.97	0.97	0.96	0.99	1.00			
#Cells SWS	0.93	-0.05	0	0.86	0.87	0.68	0.89	0.91	1.00		
#Cells FL Yellow	0.92	-0.09	-0.04	0.88	0.87	0.87	0.86	0.9	0.97	1.00	
#Cells FL Orange	0.94	-0.06	-0.02	0.89	0.88	0.89	0.9	0.92	0.97	0.96	1.00

Todos os números em negrito referem-se à significância estatística. Desta maneira, todas as variáveis selecionadas apresentam-se com algum grau de correlação. Este fato é comumente verificado em matrizes citométricas. Em nosso caso, a variável *# Cells Orange* é a que demonstra o maior nível de correlação seguida de *# Cells Yellow* e *# Cells SWS*. Porém, estas duas variáveis apresentam-se sem correlação com a variável *Avg FL Orange* e *Avg FL Yellow*. Este fato pode ser explicado devido a partículas formadas por agrupamento celulares dificultarem o acesso do fluorocromo para a marcação do ácido nucléico.

A Figura 4.9 apresenta a distribuição dos dados selecionados e suas respectivas correlações. Pode ser observado que com exceção das variáveis *Avg FL Orange* e *Avg FL Yellow* todas apresentam uma forte distribuição do tipo *skewness*. Tem-se ainda a possibilidade de identificação dos dados considerados como *outliers*. Para estes, o critério de exclusão foi o de eliminar todos os valores maiores que duas vezes o valor do desvio padrão da variável em questão.

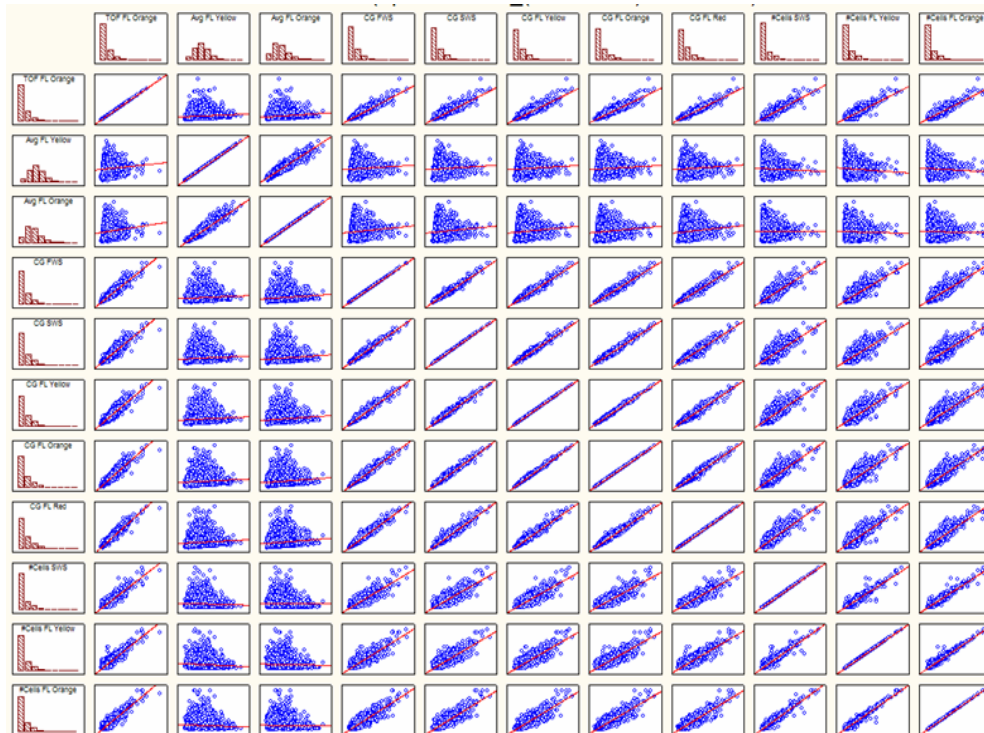


Figura 4.9 – Distribuição dos dados de acordo com as suas respectivas correlações. Esta abordagem permite a visualização de *outliers*.

A Figura 4.10 apresenta o estado final das variáveis selecionadas de acordo com o modelo de diagrama *Box & Whisker* comumente utilizados em análises estatísticas.

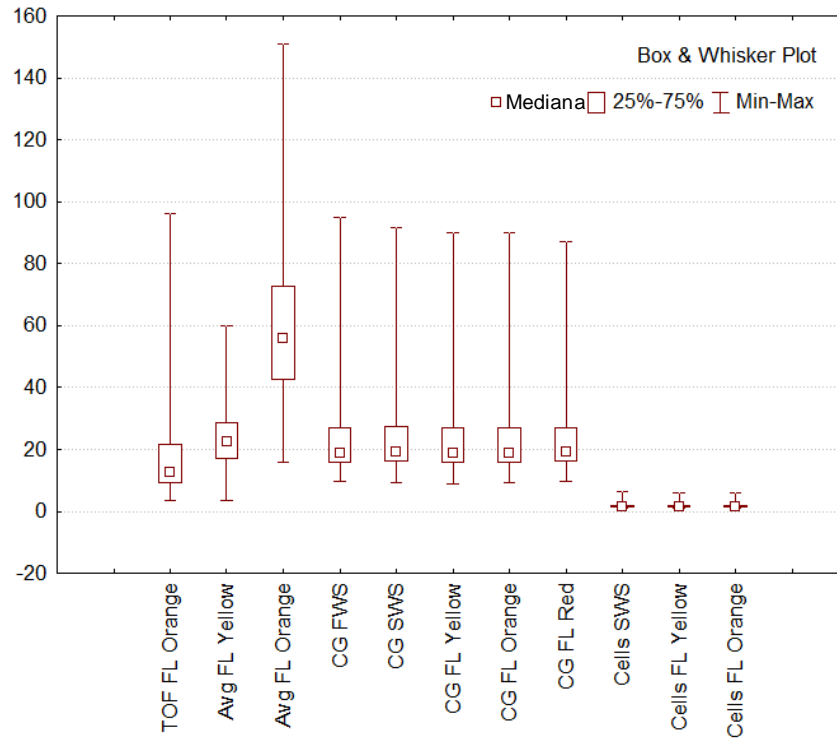


Figura 4.10 – Distribuição estatística das variáveis segundo o modelo *Box & Whisker*. Os pequenos quadrados representam as medianas, os retângulos significam a distribuição dos dados no intervalo de 25 e 75% e as barras, os valores máximos e mínimos.

Outra abordagem estatística considerada foi o de agrupamento através da análise de *Clusters* através do método de partição *K-means*. Neste sentido, a Figura 4.11 apresenta o dendrograma resultante desta abordagem.

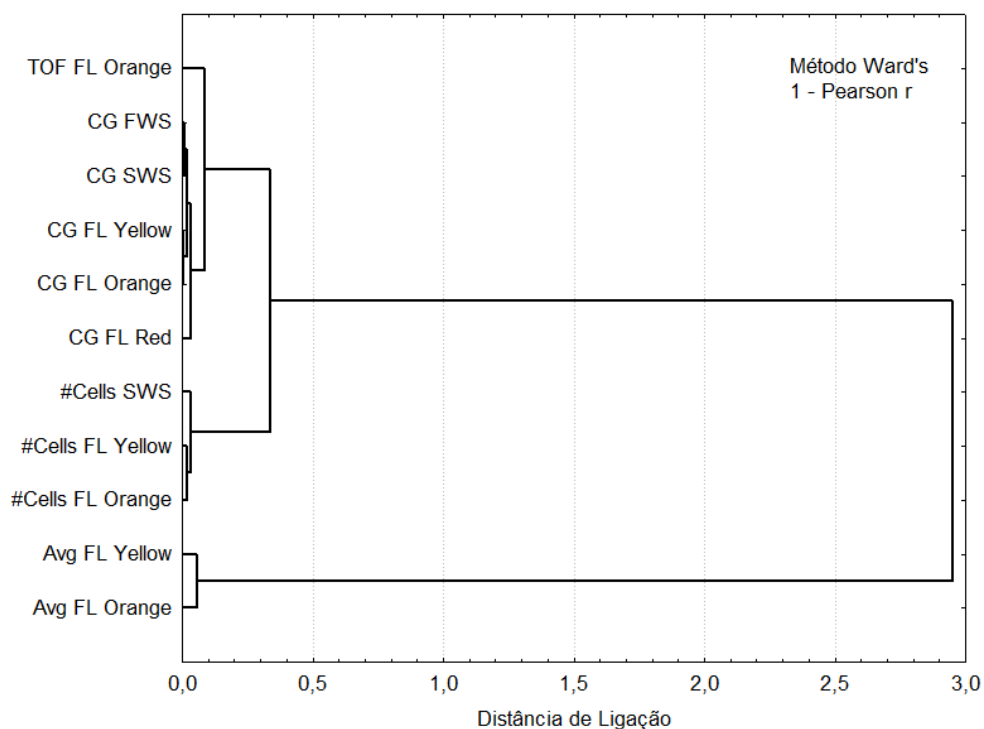


Figura 4.11 – Dendrograma produzido pelo algoritmo k-means segundo a aplicação do método Ward's e o r de Person como medida de distância.

Podem ser observados dois *clusters* na Figura 4.11, onde se verifica que um destes encontra-se também subdividido em 2 sub-grupos menores. Este padrão de distribuição de variáveis pode também ser visualizado na Figura 4.8 referente a abordagem de ACP. Este fato também foi encontrado por PEREIRA *et al.*, 2008.

Em outra abordagem aplicou-se o método *k-means* para agrupar os objetos da matriz citométrica (casos). Neste experimento foi utilizada a técnica de Ward's com a medida de distância euclidiana. A Figura 4.12 apresenta o dendrograma desta aplicação, onde são visualizados 3 grandes grupos no valor 1500 da distância de ligação.

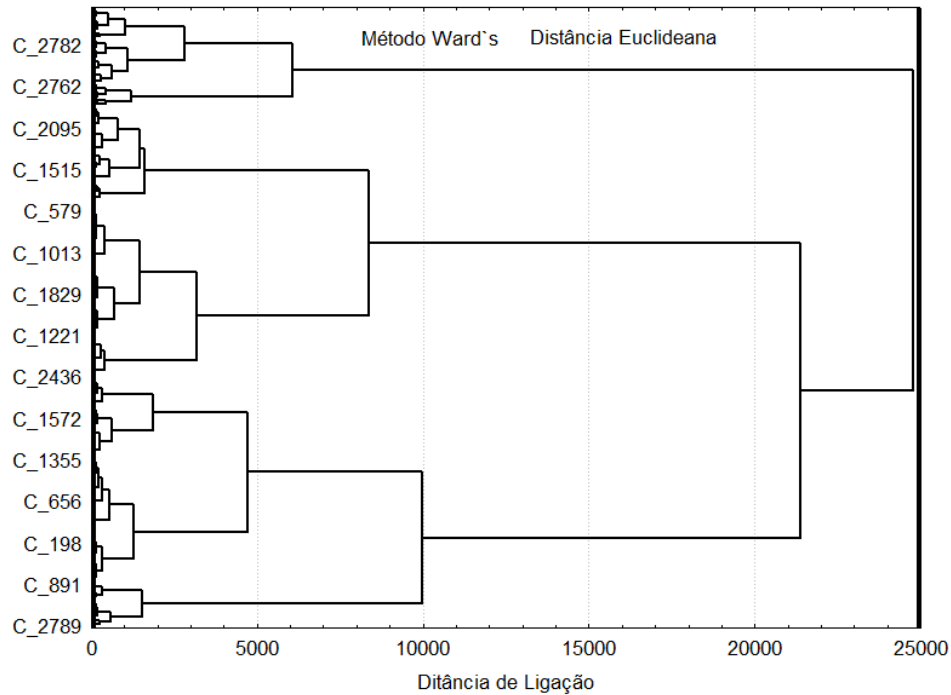


Figura 4.12 – Dendrograma produzido pela aplicação do algoritmo *K-means* com o método de Ward's e distância euclidiana.

A estatística descritiva que caracteriza cada agrupamento (*clusters*) apresenta-se na Tabela 4.5.

Tabela 4.5 – Estatística descritiva de média, desvio padrão, variância e número de casos encontrados em cada grupo.

	Cluster 1- 707 casos			Cluster 2- 1662 casos			Cluster 3- 463 casos		
	Média	Desv. Pad.	Variância	Média	Desv. Pad.	Variância	Média	Desv. Pad.	Variância
TOFFLOrOrange	15,63	8,57	73,50	13,33	8,38	70,34	34,36	15,21	231,39
AvgFLYellow	33,71	7,56	57,3	18,43	5,6	31,43	28,10	9,12	83,26
AvgFLOrOrange	87,97	19,73	389,29	45,61	12,32	152,01	71,83	23,19	538,22
CGFWS	21,4	8,53	72,84	19,81	8,08	65,44	37,98	14,42	208
CGSWS	21,84	8,67	75,21	20,18	8,4	70,59	38,32	14,32	205,28
CGFLYellw	21,25	8,6	74,01	19,62	8,19	67,16	37,74	14,26	203,4
CGFLOrOrange	21,3	8,51	72,51	19,66	8,05	64,87	37,97	14,29	204,32
CGFLRed	21,72	8,09	65,57	19,88	7,93	62,9	38,52	13,95	194,68
#Cell SWS	1,47	0,44	0,19	1,53	0,45	0,2	2,75	0,9	0,82
#Cell FLYellow	1,51	0,45	0,2	1,6	0,49	0,24	2,77	0,89	0,8
#CellsFLOrOrange	1,48	0,44	0,19	1,55	0,47	0,22	2,73	0,88	0,79

Considerando que TOF é o tempo de deformação (desvio) da luz do laser quando da passagem da partícula e a velocidade do fluxo amostral é conhecida (2ms^{-1}) e constante, esta variável reflete o tamanho da célula. Desta maneira, pode ser observado que no cluster 3 as 463 células encontravam-se em processo adiantado de divisão celular, pois seus respectivos tamanhos médios são o dobro, quando comparadas com os outros *clusters*. Este fato pode ser comprovado ainda quando se observam os valores relativos às variáveis # Cells SWS, # Cells FL Yellow e # Cells FL Orange. Uma vez que a intensidade de fluorescência é diretamente proporcional a quantidade de ácido nucléico marcado, observa-se que o *cluster* 1 (707 células) apresenta o dobro de fluorescência quando comparado com o *cluster* 2. Isto sugere que as células do *cluster* 1 encontram-se em um momento em que acabaram de efetuar a duplicação do material genético (DNA). Já o *cluster* 2, grupo com maior quantidade de células (1662 células), apresenta valores médios de marcações fluorescentes sugerindo uma menor quantidade de ácido nucléico.

Por outro lado, a busca por arquiteturas neurais do tipo MLP otimizadas pelo algoritmo genético e utilizada neste trabalho são apresentadas nas Tabelas 4.6 e 4.7.

Neste sentido, a Tabela 4.6 apresenta o resumo deste processo de otimização neural e o desempenho das três melhores redes selecionadas para os conjuntos de treinamento (70%), teste (20%) e validação (10%).

Tabela 4.6 - Resumo dos resultados do processo de otimização de arquiteturas neurais e seus respectivos desempenho.

Rede	Treinamento	Teste	Validação	Algoritmo	Função de Erro	Ativação Camada Escondida	Ativação Camada Saída
MP11-23-3	94,1088	93,0605	94,3005	BFGS43	SOS	Tang hiperbólica	Logística
MP11-30-3	92,5348	93,2383	91,81495	BFGS63	SOS	Logística	Exponencial
MP11-28-3	92,6386	92,8256	93,5931	BFGS79	SOS	Logística	Tang hiperbólica

Como pode-se observar os três modelos empregaram o algoritmo BFGS e a função de erro SOS. Apesar dos três modelos selecionados apresentarem ótimo desempenho, considera-se a rede número 1 (primeira linha da tabela) como sendo a melhor, pois apresentou numericamente o melhor desempenho, um erro de treinamento de 94,10%, teste de 93,06% e validação de 94,30%. Esta rede MLP é composta por 11

neurônios na camada de entrada, 23 neurônios na camada escondida e 3 neurônios na camada de saída. Portanto, o algoritmo genético gerou uma rede com representação do tipo *One-of-N*, utilizando uma tangente hiperbólica como função de ativação dos neurônios da camada escondida e uma função logística para os neurônios da camada de saída.

Na Tabela 4.7 apresenta-se a análise de sensibilidade das variáveis para estes três modelos de redes neurais. Pode-se observar que os maiores valores encontrados foram 4,8012 para a variável Avg FLO e 2,7904 para Cells SWS, sendo os valores das variáveis mais significativas para os modelos de redes neurais .

Tabela 4.7 – Análise de Sensibilidade das variáveis para os modelos de redes neurais.

Redes	TOF FLO	Avg FLY	Avg FLO	CG FWS	CG SWS	CG FLY	CG FLO	CG FLR	Cells SWS	Cells FLY	Cells FLO
MLP 11-23-3	1.238	1.9154	4.8012	1.5095	1.3172	1.2636	1.1162	1.1434	2.7904	1.1448	1.2552
MLP 11-30-3	1.1407	1.524	3.771	1.0744	1.0676	1.2167	1.1316	1.08	1.9559	1.1676	1.046
MLP 11-28-3	1.1349	1.5396	3.1317	1.0566	1.0286	1.0939	1.0758	1.0776	1.5412	1.1078	1.1483

A Tabela 4.8 apresenta a matriz de confusão ou matriz de erro usada para avaliar o resultado de um experimento de classificação. Os componentes da diagonal principal da matriz fornecem o número de registros ou casos classificados corretamente. Os valores nas linhas e colunas significam os erros ocorridos em cada *cluster*. Assim, o modelo 1 apresenta 391 casos classificados corretamente para o *cluster* 1, um erro designado para o cluster 2, mas que na realidade pertencente ao *cluster* 1, e 45 células classificadas erroneamente como pertencentes ao *cluster* 3. Nesta categoria (*cluster*) o modelo obteve uma acurácia de 87,86% e um erro de 12,14%. Este mesmo modelo apresentou 1151 casos classificados corretamente para o *cluster* 2 e dois casos classificados erroneamente para o *cluster* 1, e 14 como pertencentes ao *cluster* 3, portanto, uma acurácia de 98,62% e uma taxa de erro de 1,39% para esta categoria celular. Em relação ao *cluster* 3 o modelo neural obteve 311 acertos e 27 erros designados para o clusters 1 e outros 27 para o *cluster* 2. Para esta categoria a acurácia obtida foi de 85,20% e uma taxa de erro de 14,8%.

Tabela 4.8 – Matriz de Erro de treinamento dos modelos de redes neurais.

Redes		Cluster 1	Cluster 2	Cluster 3
MLP 11-23-3	Cluster 1	391	2	27
	Cluster 2	1	1151	27
	Cluster 3	45	14	311
MLP 11-30-3	Cluster 1	387	6	41
	Cluster 2	7	1148	37
	Cluster 3	43	13	287
MLP 11-28-3	Cluster 1	379	8	31
	Cluster 2	10	1143	32
	Cluster 3	48	16	302

A Tabela 4.8 também apresenta as três redes neurais do tipo MLP selecionadas, onde evidencia-se na camada de entrada as 11 variáveis, as camadas escondidas que apresentaram respectivamente 23, 30 e 28 neurônios e 3 neurônios na camada de saída.

CAPÍTULO V

Conclusões

Durante um bioprocesso é importante monitorar não somente a proliferação celular e o nível de viabilidade ou atividade metabólica mas também a formação do produto. Medidas exatas sobre a concentração de biomassa são importantes na tomada de decisão em um bioprocesso. As técnicas clássicas de monitoramento da proliferação celular apresentam várias restrições, como no caso da densidade ótica e peso seco, pois informam sobre a proliferação celular mas não contemplam o estado fisiológico das células. Neste caso, a citometria de fluxo mostra-se como uma técnica que pode fornecer estas medidas em tempo real de maneira bastante rápida e eficiente.

Neste trabalho propôs-se testar o equipamento CytoSence/Cytopsub como ferramenta de monitoramento de uma cultura bacteriana. Neste ponto específico, pode-se concluir que a abordagem proposta foi bem sucedida pois, identificou vários grupos celulares que apresentaram diferentes propriedades ópticas e citométricas, além de estabelecer a curva e as taxas de crescimento populacional ao longo do período estudado. No que se refere ao aparelho testado, o trabalho obteve êxito, porém será importante a execução de novos ensaios, comparando a performance obtida neste equipamento com os resultados de outros citômetros de capacidade de detecção semelhante e efetuar uma intercalibração como apresentado recentemente por THYSSEN *et al.*, 2008. Considerando que o CytoSub é um protótipo, e neste caso, o único capaz de produzir assinaturas ópticas na forma de *pulse shapes*, esta característica contribuiu enormemente na caracterização de determinados tipos celulares (subpopulações).

No caso de medidas de biomassa, vale ressaltar para a discussão acerca do que é exatamente este parâmetro. Usando-se os *pulse shapes* pode-se medir muito fidedignamente o tamanho e a forma geral das partículas como apresentado por TAKABAYASHI *et al.*, 2006, e aplicando-se uma simples rotação no eixo (temporal) da medida do sensor de espalhamento da luz frontal (*forward scatter*) obtém-se uma figura tridimensional cuja integral do volume é uma estimativa do chamado biovolume.

O somatório de todos os biovolumes individuais resultaria então no biovolume da população que poderia ser normalizado pelo tamanho da amostra. Por outro lado, existem autores que consideram a biomassa como sendo a quantidade de carbono incorporado, o que levaria a estratégia de monitoramento do bioprocessamento para uma visão mais química. Na realidade, até o presente momento, diversas tecnologias surgiram com o objetivo de monitorar e otimizar os biorreatores porém, sempre “esqueceram” de monitorar o mais importante, o próprio organismo que está sendo cultivado. Neste sentido a citometria de fluxo apresenta-se como uma alternativa extremamente promissora pois é capaz de efetuar uma análise de células a nível individual.

A abordagem de Mineração de Dados utilizada mostrou-se bastante efetiva principalmente na fase de pré-processamento e descrição dos dados. A utilização dos métodos estatísticos foi capaz de reduzir a dimensionalidade dos dados e assim retirar do problema as variáveis desnecessárias que somente aumentariam o custo computacional e a performance do algoritmo genético. Como apresentado por AL-HADDAD *et al.*, 2000, as redes neurais artificiais têm sido cada vez mais utilizadas na tarefa de identificação microbiológica. Neste trabalho, o modelo de rede neural apresentado foi capaz de identificar e classificar os diferentes padrões celulares e seus respectivos estados fisiológicos. Esta informação é extremamente relevante no que tange a otimização de bioprocessos pois qualquer que seja o produto objeto do bioprocessamento, este será sintetizado em alguma fase do ciclo de divisão celular. Portanto, qualquer modelo ou sistema que tenha por objetivo a otimização de bioprocessos deverá primeiramente reconhecer estas diferenças na população de células. Assim, as conclusões apresentam-se como a seguir:

- ❖ O teste do equipamento *CytoSense/CytoSub* como ferramenta de monitoramento de padrões de células em cultura pode ser considerado bem sucedido, pois demonstrou três subpopulações bacterianas em diferentes estados fisiológicos ao longo do ciclo celular,
- ❖ A *E. coli* DH10b cultivada no meio LB apresentou uma curva considerada padrão como as encontradas em culturas do tipo *batch*, diferindo pouco da chamada “curva ideal” apresentada por MADIGAN *et al.* 2004,
- ❖ A cultura demonstrou um alto grau de viabilidade celular (98%) com o emprego de CFDA,

- ❖ O fluorocromo SYBR Green I demonstrou o conteúdo de DNA nas células, porém a configuração ótica do *CytoSense/CytoSub* apresenta sensibilidade nos canais de fluorescência laranja e amarelo para o espectro de fluorescência deste fluorocromo. Este fato explica a alta correlação entre estas duas variáveis,
- ❖ O “arrasto” do número de células encontrados nos tempo T10 e T12 da Figura 4.6 sugerem a heterogeneidade e a perda da sincronia da população devido a depleção do meio.
- ❖ A capacidade de escaneamento deste citômetro de produzir assinatura ótica chama a atenção para a possibilidade de se efetuar medidas de tamanho, forma e biomassa (biovolume) das partículas individualmente. Em nosso caso, a integral da curva apresentada pela variável TOF FLO seria a estimativa do biovolume (conteúdo de DNA) da célula. A mesma abordagem pode ser utilizada para medir a produção de uma enzima através do uso de *primers* fluorescentes específicos.
- ❖ O algoritmo genético mostrou-se bastante efetivo no processo de otimização da arquitetura de modelos de redes neurais de alto desempenho, envolvidas na tarefa de classificação de padrões.

A citometria de fluxo é uma tecnologia que vem evoluindo rapidamente e promovendo um grande avanço, principalmente quando fluorocromos específicos são acoplados as técnicas de biologia molecular como na utilização de anticorpos ou *primers* específicos e a possibilidade de hibridização *in situ* em citômetros multiparamétricos e equipados com mais de um laser.

Como perspectivas futuras pretende-se:

- ❖ Adquirir experiência em clonagem de inserção de plasmídeos,
- ❖ efetuar a técnica de hibridização *in situ* acoplada a citometria de fluxo,
- ❖ utilizar *primers* fluorescentes específicos,
- ❖ desenvolver um modelo de conhecimento baseado em Regras de Associação, que correlacionem fatores físicos e químicos da cultura com medidas biológicas da expressão gênica e seu produto, através de citometria de fluxo.

REFERÊNCIAS BIBLIOGRÁFICAS

- ABU-ABSI, R. N., ZAMAMIRI, A., KACMAR, J., BALOGH, S. J., SRIENC, F., 2003, “Automated flow cytometry for acquisition of time-dependent population data”, *Cytometry*, v. 51, pp. 87-96.
- AL-HADDAD, L., MORRIS, C. W., BODDY, L., 2000, “Training radial basis function neural networks: effects of training set size and imbalanced training sets”, *Journal of Microbiological Methods*, v. 43, pp. 33-44.
- ANDREW, S. & BAILEY, M.J., 2000, “Bacterial community structure and physiological state within an industrial phenol bioremediation system”, *Appl. Envir. Microb.*, v.66, pp. 2400-2407.
- ARMITAGE, J. P., ROLLAND, I. B., JENAL, U., KENNY, B., 2005, “Neural networks in bacterial: making connections”. *Journal of Bacteriology*, v. 187, pp. 26-36.
- ASSIS, A. J. & MACIEL FILHO, R., 2000, “Soft sensors development for on-line bioreactor state estimation”, *Computers and Chemical Engineering*, v. 24, pp. 1099-1103.
- BALÁZSI, G., BARABÁSI, A. L., OLTAVAI, Z. N., 2005, “Topological units of environmental signal processing in the transcriptional regulatory network of *Eschericia coli*”. In: *Proceedings of National Academy of Science*, v. 102, n. 22, pp. 7841-7846.
- BARBU, M., CARAMAN, S., CEANGĂ, E., 2005, “Bioprocess control using a recurrent neural network model”, In: *Proceedings of the 2005 IEEE: International Symposium on Intelligent Control*, pp. 479-484, Limassol, June.
- BASSLER, B. L., 1999, “How bacteria talk to each other: regulation of gene expression by quorum sensing”, *Current Opinion in Microbiology*, v. 2, pp. 852-887.
- BERTONI, G., DEHÒ, G., 2001, “Bacteriophage P2 recombination in the superinfection preprophage state and under replication control by phage P4”, *Mol. Gen. Genet.*, v. 266, pp. 406-416.

- BHOWMIK, G., SAHA, G., BARUA, A., SINHA, S., 2000, "On-line detection of contamination in a bioprocess using artificial neural networks", *Chemical Engineering & Technology*, v. 23, n. 6, pp. 543-549.
- BICIATO, S., PANDIN, M., DIDONÈ, G., DI BELLO, C., 2002, "Pattern identification and classification in gene expression data using an autoassociative neural network model". *Biotechnology and Bioengineering*, v. 81, n. 5, pp. 594-606.
- BIRAN, I. & WALT, D. R., 2002, "Optimal imaging fiber-based single live cell arrays a high-density cell assay platform", *Anal. Chem.*, v. 74, pp. 3046-3054.
- BOUVIER, T., TROUSSELLIER, M., ANZIL, A. COURTIES, C., SERVAIS, P., 2001, "Using light scatter signal to estimate bacterial biovolume by flow cytometry", *Cytometry*, v. 44, pp.188-194.
- BREHM-STECHER, B. F. & JOHNSON, E. A., 2004, "Single-cell microbiology: tolls technologies and applications", *Microbiol. Mol. Biol. Rev.*, v. 68, pp. 538-559.
- BROWN, M. R. W., COLLIER, P. J., GILBERT, P., 1990, "Influence of grown rate on susceptibility to antimicrobial agents: modification of the cell developpe and batch and continius culture studies", *Antimicrob. Agents Chemother*, v. 34, pp. 1623-1628.
- BUNTEMEYER, H., MARZAHN, R., LEHMANN, J., 1994, "A direct computer control concept for mammalian cell fermentation processes", *Cytotechnology*, v. 15, pp. 271-279.
- BUSAM, S., MCNABB, M., WACKWITZ, A., SENEVIRATHNA, W., BEGGAN, S., VAN DER MEER, J. R., WELLS, M., BREUER, U., HARMAS, H., 2007, "Artificial neural network study of whole-cell bacterial bioreporter response determined using fluorescence flow cytometry", *Anal. Chem.*, v. 79, n. 23, pp. 9107-9114.
- CÁNOVAS, M., GARCÍA, V., BERNAL, V., TORROGLOSA, T., IBORRA, J. L., 2007, "Analysis of Escherichia coli cell state by flow cytometry during whole cell catalyzed biotransformation for L-carnitine production", *Process Biochemistry*, v.42, pp.25-33.

- CIMANDER, C., BACHINGER, T., MANDENIUS, C-F., 2003, "Integration of distributed multi-analyzer monitoring and control in bioprocessing based on a real-time expert system", *Journal of Biotechnology*, v. 103, pp. 237-248.
- CLEMENTSCHITSCH, F. & BAYER, K., 2006, "Improvements of bioprocess monitoring: development of novel concepts", *Microbial Cell Factories*, v. 5, n.19, pp. 1-11.
- COSTELLO, E. K., LAUBER, C. L., HAMADY, M., FIERER, N., GORDON, J. I., KNIGHT, R., 2009, "Bacterial community variation in human body habitats across space and time", *Science*, v.326, pp. 1694-1697.
- Current Protocols of Cytometry* (2006) da *International Society of Analytical Cytology* – ISAC
- DAVEY, H. M. & KEEL, D. B., 1996, "Flow cytometry and cell sorting of heterogeneous microbial", *Microbiol. Rev.*, v. 60, pp. 641-696.
- DE VEAUX, R. D., BAIN, R., UNGAR, L. H., 1999, "Hybrid neural network models for environmental process control", *Environmetrics*, v.10, n. 3, pp.225-236.
- DIAZ, C., DIEU, P., FEUILLERAT, C., LELONG, P., SALOMÉ, 1995, "Adaptive predictive control of dissolved oxygen concentration in a laboratory-scale bioreactor", *Journal of Biotechnology*, v. 43, pp.21-32.
- DOOLITTLE, W. F., 1999, "Phylogenetic classification and the universal tree", *Science*, v. 284, pp.2124-2129.
- DUBELAAR, G. B. J., GERRITZEN, P. L., BEEKER, A. E. R., JONKER, R. R., TANGEN, K., 1999, "Design and first results of CytoBuoy a wireless flow cytometer for in situ analysis of marine and fresh waters", *Cytometry*, v. 37, pp.247-254.
- DUBELAAR, G. B. J. & GERRITZEN, P. L., 2000, "CytoBuoy: a step forward towards using flow cytometry in operational oceanography", *Sci. Mar.*, v.64, n. 2, pp.255-265.
- DUBELAAR, G. B. J., VENKAMP, R. R., GERRITZEN, P. L., 2003, "Handsfree counting and classification of living cells and colonies" In: 6th Congress on Marine Sciences, Havana.

- DURFEE, T., NELSON, R., BALDWIN, S., PLUNKETT III, G., BURLAND, V., MAU, B., PETERSINO, J. F., QIN, X., MUZNY, D. M., AYELE, M., GIBBS, R. A., CSÖRGŐ, B., PÓSFAL, G., WEINSTOCK, G. M., BLATTNER, F., 2008, "The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhouse", *Journal of Bacteriology*, v.190, pp. 2597-2606.
- DÜRRSCHMID, E.; SPANNBAUER, B.; STRIEDNER, G.; CLEMENTSCHITSCH, F., BAYER, K., 1998, "Optimized data exploration recombinant fermentation using neural network simulations", *7th Intern. Conference on Computer Application in Biotechnology*, Osaka, May 31-June 4, Japan.
- ENGELBERG-KULKA, H., SAT, B., RECHES, M., AMITAI, S., HAZAN, R., 2003, "Bacterial programmed cell death systems as targets for antibiotics", *Trends Microbiol.*, v. 12, pp. 66-71.
- EVERITT, B. S., 1993, *Cluster Analysis*. 3 ed., Halsted Press.
- FAYYAD, U. M., SHAPIRO, P., SMYTH, G. P., UTHURUSAMY, R., "Advances in Knowledge Discovery and Data Mining", In: AAAI Press, The MIT Press, 1996.
- FETECAU, G., NICOLAU, V., PALADE, V., FETECAU, M., 2003, "Intelligent optimal control of a biosynthesis process using a neural network based estimator". In: *Proceedings of Kes*, pp. 941-949.
- FUQUA, C. & GREENBERG, E. P., 2002, "Listening in on bacteria: acyl-homoserine lactone signaling", *Nature*, v. 3, pp. 595-685.
- GAJKOWSKA, A., OLDAK, T., JASTRZEWSKA, M., MACHAJ, E. K., WALEWASKI, J., KRASZEWSKA, E., POJDA, Z., 2006, "Flow cytometric enumeration of CD34+ hematopoietic stem and progenitor cells in leukapheresis product and bone marrow for clinical transplantation: a comparison of three methods", *Folia Histochemica et Cytobiologica*, v.44, pp. 53-60.
- GLASSEY, J., IGNOVA, M., WARD, A. C., MONTAGUE, G. A., MORRIS, A. J., 1997, "Bioprocess supervision: neural networks and knowledge based systems", *Journal of Biotechnology*, v. 52, pp. 201-205.

- GONÇALO, M. Análise em Componentes Principais, ISEL, 2004.
<http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pes/trab3.pdf> último acesso em 14/12/2008.
- HAN, J. & KAMBER, M., 2001, *Data Mining: concepts and techniques*. 1 ed. San Diego, Academic Press.
- HAYKIN, S., 2001, *Redes Neurais, princípios e prática*, 2 ed., Rio de Janeiro, Bookman.
- HEWITT, C. J., NEBE-VON-CARON, G., NIENOW, A. W., 2000, “Use of multi-staining flow cytometry to characterize the physiological state of *Escherichia coli* W3110 in high cell density fed-batch cultures”, *Biotechnol. Bioeng.*, v. 63, pp. 705-711.
- HEWITT, C. J. & NEBE-VON-CARON, G., 2001, “An industrial application Assessment of cell physiological state and its application to the study of microbial fermentation”, *Cytometry*, v.44, pp. 179-187.
- HOLLAND, J. H., 1975, *Adaptation in natural and artificial systems*, MIT Press.
- HOLT, J. G., KRIEG, N. R., SNEATH, P.H.A., STALEY, J. T., WILLIAMS, S.T., 1994, *Bergey's manual of determinative bacteriology*, 9 ed., Maryland, Williams & Wilkis.
- HRISTOVA, K & PATARINSKA, T., 2006, “Neural network modelling of continuous microbial cultivation accounting for the memory effects”. *International Journal of Systems Science*, v. 37, pp. 271-277.
- JAYALAKSHIMI, G, A., MUTHARASU, D., RAJARAM, S., PANDIYAN, S. G., KANNIAPPAN, P., 2000, “An Evolutionary Programming Approach to Evolve the Architecture of Artificial Neural Networks”.
<http://www.ise.nus.edu.sg/proceedings/apros2000/fullpapers/28-06.htm>
- JENZSCH, M., GNOTH, S., KLEINSCHMIDT, M., 2006, “Improving the batch-to-batch reproducibility in microbial cultures during recombinant protein production by guiding the process along a predefined total biomass profile”, *Bioprocess Biosyst. Eng.*, v.29, pp. 315-321.
- KARIM, M. N. & RIVERS, S., 1992, “Artificial neural networks in bioprocess state estimation”, *Adv. In Biochem. Eng. And Biotechnol*, v.46, pp. 2-33.

- KIKUCHI, S., TOMINAGA, D., ARITA, M., TAKARHASHI, K., TOMITA, M., 2003, "Dynamic modeling of genetic networks using genetic algorithm and S-system". *Bioinformatics*, v. 19, pp. 643-650.
- KONEMAM, E. W., ALLEM, S. D., SCHREKENBERGER, P., JANDA, C., WINN, W. C., 1997, **Color atlas and textbook of diagnostic microbiology**, 5 ed. Philadelphia, Lippincott company.
- KONG, D., GENTZ, R., ZHANG, J., 1998, "Development of a versatile computer integrated control system for bioprocess controls", *Cytotechnology*, v. 26, pp. 227-236.
- KRSTIC, V., MAGLICA, Z., PALJETAK, H. C., PODOBNIK, B., PAVIN, N., 2007, "Min-protein oscillation in *E. coli*: three-dimensional off-lattice stochastic reaction diffusion model", *Journal of Statistical Physics*, v. 128, pp. 5-20.
- LACKNER, L. L., RASKIN, D. M., DE BOER, P. A. J., 2003, "ATP-dependent interactions between *Escherichia coli* Min proteins and the phospholipid membrane in vitro", *Journal of Bacteriol.*, v. 185, n. 3, pp. 735-749.
- LAUB, M. T., MACADAMS, H. H., FELDBLYUM, T., FRASER, C.M., SHAPIRO, L., 2000, "Global analysis of the genetic network controlling a bacterial cell cycle", *Science*, v. 290, pp. 2144-2148.
- LIAO, J. C., BOSCOLO, R., YANG, Y-L, TRAN, L. M., SABATTI, C, ROYCHOWDHURY, V. P., 2003, "Network component analysis: reconstruction of regulatory signals in biological systems", *PNAS*, v, 100, pp. 115524-115527.
- LOSSER, V., HAMMES, F., KELLER, M., BERNEY, M., KOVAR, K., EGLI, T., 2005, "Flow-cytometric detection of changes in the physiological state of *E. coli* expressing a heterologous membrane protein during carbon-limited fedbatch cultivation", *Biotechnology and Bioengineering*, v. 92, n. 1, pp. 69-78.
- MADIGAN, M. T., MARTINKO, J. M., PARKER, J., 2004, **Microbiologia de Brock**. 10 ed. São Paulo, Prentice Hall.
- MANDENIUS, C-F., 2004, "Recent developments in the monitoring, modeling and control of biological production systems", *Bioprocess Biosyst. Eng.*, v. 26, pp. 347-351.

- MARTINEZ, A. A., COLLADO-VIDES, J., 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, v. 6, pp. 482-489.
- McADAMS, H. H., SHAPIRO, L., 2003, “A bacterial cell-cycle regulatory network operating in time and space”, *Science*, v. 301, pp. 1874-1877.
- MITCHELL, D. A., VON MEIEN, O. F., KRIEGER, N., DALSENTER, F. D. H., 2004, “A review of recent developments in modeling of microbial growth kinetics and intraparticle phenomena in solid-state fermentation”, *Biochem. Engin. Journal*, v. 17, pp. 15-26.
- MÜLLER, S., 2007, “Modes of cytometric bacterial DNA pattern: a tool for pursuing growth”, *Cell Prolif.*, v. 40, pp. 621-639.
- NA, J. G., CHANG, Y. K., CHUNG, B. H., LIM, H. C., 2002, “Adaptive optimization of fed-batch culture of yeast by using genetic algorithms”, *Bioprocess and Biosystems Engin.*, v. 24, pp. 299-308.
- NEBE-VON-CARON, G, STEPHENS, P., BADLEY, A. R., 1999, “Bacterial detection and differentiation by cytometry and fluorescent probes”, *Royal Microbiology Society*, v.34, n.1., pp. 321-327.
- NEBE-VON-CARON, G., STEPHENS, P. J., HEWITT, C. J., POWELL, J. R., BADLEY, R. A., 2000, “Analysis of bacterial function by multi-color fluorescence flow cytometry and single sorting”, *J. Microbiol. Methods*, v. 42, pp. 97-114.
- NING, S. B., GUO, H. L., WANG, L. SONG, Y. C., 20002, “Salt stress induces programmed cell death in prokaryotic”, *J. Appl. Microbiol.*, v. 93, pp. 15-28.
- NUNEZ, R., 2001, “DNA measurement and cell cycle analysis by flow cytometry”, *Curr. Issues Mol .Biol.*, v.3, pp. 67-70.
- OCHAMAN, H., LAWRENCE, J. G., GROISMAN, E. A., 2000, “Lateral gene transfer and the nature of bacterial innovation”, *Nature*, v. 405, pp. 299-304.
- PATNAIK, P. R., 2007, “Intelligent descriptions of microbial kinetics in finitely sispered bioreactors: neural and cybernetic models for PHB biosynthesis by *Ralstonia eutropha*”, *Microbial Cell Factories*, v. 6, n. 23, pp. 1-8.

- PEREIRA, G. C., 2005, *Mineração de dados para análise e diagnóstico ambiental*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, Brasil.
- PEREIRA, G. C., COUTINHO, R., EBECKEN, N. F. F., 2008, ‘Data Ming for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo’, *Brazilian Journal of Oceanography*, v.56, n.1, pp.1-12.
- PIGRAM, G. M. & MACDONALD, T. R., 2001, “Use of neural network models to predict industrial bioreactor effluent quality”, *Environ. Sci. Technol.*, v. 35, pp. 157-162.
- POPOVA, S., KOPRINKOVA, P., PATARINSKA, T., 2003, “Neural network based biomass and growth rate estimation aimed to control of a chemostat microbial cultivation”, *Appl. Artificial Intelligence*, v. 17, n. 4, pp. 345-360.
- PORTER, J., ROBINSON, J., D., PICKUP, R., EDWARDS, C., 1995a, “Recovery of a bacterial sub-population from swage using immunofluorescence flow cytometry and cell sorting”, *FEMS Microb. Lett.*, v.133, pp. 195-199.
- PORTER, J., EDWARDS, C., PICKUP, R., 1995b, “Rapid assessment of physiological status in *Escherichia coli* using fluorescent probes”, *Journal Appl. Bacteriology*, v.79, pp. 399-408.
- PORTER, J., DEERE, D., HARDMAN, M., EDWARDS, C., PICKUP, R., 1997, “Go with the flow – use of flow cytometry in environmental microbiology”, *FEMS Microb. Ecology*, v.24, pp. 93-101.
- PRICE, M. H., HUANG, K. H., ALM, E. J., ARKIN, A. P., 2005a, “A novel method for accurate operon prediction in all sequence prokaryotic”, *Nucleic Acids Res.*, v. 33, pp.880-892.
- PRICE, M. H., HUANG, K. H., ALM, E. J., ARKIN, A. P., 2005b, “Operon formation is driven by co-regulation and not by horizontal gene transfer”, *Genome Res.*, v. 15, pp. 809-819.
- PRICE M. N., ARKIN, A. P., ALM, E. J., 2006, “The life-cycle of operons”, *Plos Genetics*, v. 2, n. 6, pp. 859-873.
- RAINEY, P. B., BUCKLING, A., KASSEN, R., TRAVISANO, M., 2000, “The emergence and maintenance of diversity: insights from experimental bacterial populations”, *Tree*, v. 15, n.6, pp. 243-247.

- RAJWA, B., VENTKATAPATHI, M., RAGHEB, K., BANADA.P. P.,HIRLEMAN, E. D., LARRY, T., ROBINSON, J. P., 2008. “Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier”, *Cytometry*, v. 73, pp. 369-379.
- RICE, K. & BAYLES, K.W., 2003, “Death’s toolbox: examining the molecular components of bacterial programmed cell death”, *Mol. Microbiol.*, v.50, pp. 729-738.
- RICHARD, A. J & WICHERN, D. A., 1992, *Applied Multivariate Statistical Analysis*. 3 ed., New Jersey, Prentice Hall.
- RIESENBERG, D. & GUTHKE, R., 2001, “High-cell-density cultivation of microorganisms”, *Appl. Microbio. Biotechnol.*, v. 51, pp. 422-430.
- ROCA, E., FLORES, J., RODRIGUEZ, I., CAMASELLE, C., NÚÑEZ, M. J., LEMA, J. M., 2004, “Knowledge based control applied to fixed bed pulsed bioreactors”, *Bioprocess and Biosystems Engin.*, v. 14, n. 3, pp. 113-118.
- ROWICKA, M., KUDLICKI, A., TU, B. P., OTWINOWSKI, Z., 2007, “High-resolution timing of cell cycle-regulated gene expression”, *Proceedings of National Academy of Science*, v. 104, n. 43, pp. 16892-16897.
- SANTIAGO, F. E., WHITTMAN, T. S., WINKWORTH, C. L., RILEY, M. A., LENSKI, R. E., 2005, “Genomic divergence of Escherichia coli strains: evidence for horizontal transfer and variation in mutation rates”, *International Microbiology*, v.8, pp. 271-278.
- SENGUPTA, S. & RUTENBERG, A., 2007, “Modeling partitioning of Min proteins between daughter cells after septation in *Escherichia coli*”, *Phys. Biol.*, v. 4, pp. 145-153.
- SHAPIRO, H. M., 2003, *Practical flow cytometry*, 4 ed. New Jersey, John Wiley & Sons.
- SHAW, A. D., WINSON, M. K., WOODWARD, A. M., 2000, “Bioanalysis and Bioprocess Monitoring”. In: *Springer Berlin/Heidelberg*, Chapter: pp. 83-113.
- SILVA, T. L., REIS, A., HEWITT, C., ROSEIRO, J. C., 2004 Citometria de Fluxo - Funcionalidade Celular on-line em bioprocessos <http://dequim.ist.utl.pt/bbio/77/pdf/citometria2.pdf> último acesso em 14/12/2008.

- SILVA, R. G., CRUZ, A. J. G., HOKKA, C. O., GIORDANO, R. L. C., GIORDANO, R. C., 2000, "A hybrid feedforward neural network model for the cephalosporin C production process", *Brazilian Journal of Chem. Engin.*, v. 17. Doi: 10.1590/S0104-66322000000400023.
- SIMON, L. & KARIM, M. N., 2002, "Modeling and control of amino acids starvation-induced apoptosis in CHO cell cultures", *In: Proceedings of the American Control Conference*, Anchorage, AK May 8-10, pp. 1579-1584.
- SKARSTAD, K., STEEN, H. B., BOYE, E., 1983, "Cell cycle parameters of slowly growing *Escherichia coli* B/r studied by flow cytometry", *Journal of Bacteriology*, v.154, n.2, pp.656-662.
- SOGIN, M. L., MORRISON, H. G., HUBER, J.A., WELCH, D. M., HUSE, M. S., NEAL, P. R., 2006, "Microbial diversity in the deep sea and the underexplored rare biosphere". *PNAS*, v.103, pp. 12115-12120.
- SOUTOURINA, O. A., SEMENOVA, E. A., PARFENOVA, V. V., DANCHIN, A., BERTIN, P., 2001, "Control of bacterial motility by environmental factor in polar flagellated and peritrichous bacteria isolated from lake Baikal", *Appl. and Envir. Microbiol.*, v.67, n.9, pp. 3852-3859.
- STATSOFT, 2008. STATISTICA FOR WINDOWS, Statsoft. France.
- SUNRAY, M., ZURGIL, N., SHAFRAN, Y., DEUTSCH, M., 2002, "Determination of individual cell Michaelis-Menten constants", *Cytometry*, v.47, pp. 8-16.
- TAKABAYASHI, M., LEW, K., JOHNSON, A., MARCHI, A. L., DUGDALE, R., WILKERSON, F. P., 2006. "The effect of nutrient availability and temperature on chain length of the diatom, *Skeletonema costatum*" *Journal of Plankton Research*, v.28, n.9, pp.831-840.
- TEGEL, H., HEDHAMMAR, M., UHLÉN, M., OTTOSSON, J., HOBBER, S., 2007, "Flow cytometry-based analysis of promoter effects on solubility of recombinantly expressed proteins", *Journal of Biotechnology*, v. 131, n. 2, suppl.1, pp. 9-10.
- THYSSEN, M., TARRAN, G. A., ZUBKOV, M., HOLLAND, R. J., GRÉGORI, G., BURKILL, P. H., DENIS, M. 2008. The emergence of automated high-frequency flow cytometry: revealing temporal and spatial phytoplankton variability. *Journal of Plankton Research* 30(3): 333-343.

- THURSTONE, L. L., 1947, **Multiple-factor analysis**, Chicago, University of Chicago Press.
- TRABULSI, L. R., ALTERTHUM, F., GOMPERTZ, O. F., CANDEIAS, J. A. N., 2002, *Microbiologia*, 3 ed., Rio de Janeiro, Editora Atheneu.
- VINTERBO, S. A., KIM, EUN-YOUNG, OHONO-MACHADO, L., 2005, “Small, fuzzy and interpretable gene expression based classifiers”, *Bioinformatics*, v. 21, n. 9, pp. 1964-1970.
- VOHRADSKY, J., 2001, “Neural network model of gene expression”. *Faseb Journal*, v. 15, pp. 846-854.
- WITHERS, H., SWIFT, S., WILLIAMS, P., 2001, “Quorum sensing as an integral component of gene regulatory networks in Gram-negative bacteria”, *Current Opinion in Microbiology*, v. 4, pp. 186-193.
- WOLF, Y., ROGOZIN, I. B., KONDRASHOV, A. S., KOONIN, E. V., 2001, “Genome alignment, evolution of prokaryotic genome organization, and prediction of genes function using genomic context, *Genome Res.*, v.11, pp.356-372.
- YING, LI, ZENG-RONG, L., JIAN-BAO, Z., 2007, “Dynamics of network motifs in genetic regulatory networks”, *Chinese Phys.*, v. 16, n.9, pp. 2587-2594.
- ZHAO, H., GUIVER, J., NEELAKANIAN, R., BIEGLER, L. T., 2001, “A nonlinear industrial model predictive controller using integrated (PLS) and neural net state-space model”, *Contr. Eng. Prac.*, v.9, pp. 125-133.