SELEÇÃO DE GENES PREDITIVOS POR MODELO EVOLUTIVO E RECURSIVO EM MICROARRANJOS DE DNA DE MALIGNIDADES HUMANAS.

Marcelo Ribeiro-Alves

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Aprovada por:

_____

Prof. Flávio Fonseca Nobre, PhD.

_____

Prof. Márcio Nogueira de Souza, D.Sc.

_____

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

_____

Prof. Ulisses Gazos Lopes, D.Sc.

_____

Prof. André Carlos Ponce de Leon Ferreira de Carvalho, PhD.

RIO DE JANEIRO, RJ - BRASIL
AGOSTO DE 2007

RIBEIRO-ALVES, MARCELO

Seleção de Genes Preditivos por Modelo Evolutivo e Recursivo em Microarranjos de DNA de Malignidades Humanas [Rio de Janeiro] 2007

XV, 171 p. 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia Biomédica, 2007)

Tese – Universidade Federal do Rio de Janeiro, COPPE

1 - Microarranjos de DNA

2 - Seleção de Genes

3 - Algoritmos Genéticos

4 - Máquinas de Vetores de Suporte

I. COPPE/UFRJ II. Título (série)

Em memória de meu pai, Hélcio Ribeiro Alves (1941–2002).

Agradecimentos

Aos meus pais, Hélcio e Elizabeth, pelo amor incondicional e por acreditarem que a educação de um filho é sua maior riqueza e, aos meus irmãos, Hélcio Jr. e Patrícia, pelo carinho e por me darem apoio irrestrito e incondicional.

A minha esposa e companheira do dia-a-dia, Lidia, e ao nosso filho Felipe, pelo amor e carinho que me fazem acreditar que o propósito da vida não deve ser mesquinho.

Ao professor e amigo Flávio Nobre não apenas pela orientação essencial para o desenvolvimento dessa tese, de seu projeto à revisão do texto, mas principalmente pela amizade e apoio presentes.

Aos amigos e colegas do Laboratório de Eng. de Sistemas de Saúde por tornarem os dias de trabalho mais suaves, pelas caronas e pelos carinho e confiança.

Aos professores e funcionários do Programa de Eng. Biomédica e, em geral, da COPPE, que possibilitaram, apesar das pedras no caminho, que pudesse ter um curso proveitoso e produtivo.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) pela bolsa de estudos e financiamento da tese.

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

SELEÇÃO DE GENES PREDITIVOS POR MODELO EVOLUTIVO E RECURSIVO EM MICROARRANJOS DE DNA DE MALIGNIDADES HUMANAS.

Marcelo Ribeiro-Alves

Agosto/2007

Orientador: Flávio Fonseca Nobre
Programa: Engenharia Biomédica

Microarranjos de DNA permitiram a monitoração simultânea do nível de expressão de centenas de genes. Esforços foram feitos para a classificação no câncer baseada nos perfis de expressão gênica, o que envolve a identificação de subgrupos de genes preditores. Esta pode ser considerada como um problema de otimização com minimização do subconjunto de genes mantendo a acurácia de classificação. As maiores dificuldades na otimização são a esparsidade dos dados e a grande dimensão do espaço de soluções. Soluções para o problema incluem o emprego de Máquinas de Vetores de Suporte (SVMs) e Algoritmos Genéticos (GAs). As SVMs são classificadores binários capazes de encontrar margens máximas entre classes com boa generalização, enquanto os GAs mantém populações de soluções permitindo uma busca paralela eficiente em espaços grandes e complexos. Propomos o emprego de Algoritmos Genéticos Multi-Objetivos (MOGA), que minimizam o conjunto de genes, o risco empírico baseado no erro de validação cruzada *leave-one-out* e a diferença absoluta da razão de erro entre classes, enquanto escolhe o parâmetro de margem *soft*. O algoritmo proposto inclui o pós-processamento por Eliminação Recursiva de Atributos (RFE) empregando genes não-inferiores repetidos nas 10 simulações do SVM–MOGA. O algoritmo RFE–SVM–MOGA desenvolvido foi aplicado em três conjuntos de dados de domínio público: Leucemia, Linfoma e Câncer de Cólon. Foi possível classificar corretamente 100% das amostras de treino e, respectivamente, $97,05\%$, $90,90\%$ e $77,27\%$ das amostras de teste com subconjuntos gênicos de tamanho 15, 16 e 14, com baixo custo computacional, critério explícito do uso de genes selecionados em diferentes simulações do MOGA e margem de decisão linear.

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

SELECTION OF PREDICTIVE GENES BY EVOLUTIONARY AND RECURSIVE MODELS FROM DNA MICROARRAYS IN HUMAN CANCER.

Marcelo Ribeiro-Alves

August/2007

Advisor: Flávio Fonseca Nobre
Department: Biomedical Engineering

DNA microarray has the ability for measuring in parallel the expression level of thousands of genes. Several endeavors have been made for cancer classification based on gene expression profiles, and these involve the identification of subsets of predictive genes. Gene selection can be considered as an optimization problem of minimizing gene subset size while achieving reliable and accurate classification. The main difficulties in solving it are the sparsity of the data and the large search space of solutions. Approaches to address this problem include using Support Vector Machines (SVMs) and Genetic Algorithms (GAs). SVMs are binary classifiers capable to find maximal margins between classes with good generalization ability, while GAs are approaches that maintain populations of solutions allowing efficient parallel searches in large, complex and multi-modal spaces. We propose to apply a Multi-Objective Genetic Algorithm (MOGA), which minimizes the gene subset size, the empirical risk based on the leave-one-out cross-validation error, and the absolute difference in error rate between classes, while choosing the soft margin parameter. The proposed algorithm includes a Recursive Feature Elimination (RFE) post-processing step using replicated non inferior genes from 10 runs of SVM–MOGA. The developed RFE–SVM–MOGA algorithm was applied to three public available data sets: Leukemia, Lymphoma, and Colon Cancer. It was able to correctly classify 100% of the training samples and 97.05%, 90.90% e 77.27% of the test samples with gene subsets of size 15, 16 e 14, respectively, with low computational cost, explicitly criteria in the use of genes selected in different runs of MOGA, and a linear decision margin.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A *genômica do câncer* é uma busca entre indivíduos, ou agrupamentos familiares, acometidos por malignidades, afim de se obter a coleção completa de genes e mutações (tanto herdadas como somáticas), que contribuem para o desenvolvimento da célula cancerosa, assim como sua progressão de uma forma de câncer localizada para um que cresce desordenadamente e gera metástases [3].

A maioria dos cânceres surgem de várias mutações genéticas, que se acumulam nas células do corpo durante a vida. Estas envolvem genes localizados com maior freqüência nos autossomos (mutações somáticas), podendo também envolver genes localizados nos cromossomos sexuais de células de linhagem germinativa (mutações herdadas). As mutações somáticas não são passadas à geração seguinte. Todos os tumores são clonais, significando que são originados de uma única célula ancestral [4]. Uma vida de 80 anos sem câncer, por exemplo, requer que 10 bilhões de milhões de células copiem a si mesmas corretamente, ou seja, sem a presença de erros. Estas mudanças ocorrem durante o tempo de vida de uma pessoa pela exposição a carcinógenos e outros matagênicos, ou pelo erro aleatório que ocorre rotineiramente nos crescimento e divisão celulares. Ocasionalmente, uma dessas mutações somáticas altera o funcionamento de algum gene crítico, propiciando uma vantagem de crescimento para a célula onde ocorreu. Um clone então surge dessa célula única [4].

O câncer não é considerado como uma doença hereditária, porque a maioria dos casos de câncer, cerca de 80 a 90%, ocorre em pessoas sem histórico familiar da doença [5, 6]. Entretanto, a chance de uma pessoa desenvolver câncer pode ser influenciada pela herança de certos tipos de alterações genéticas. Estas alterações tendem a aumentar

a suscetibilidade individual de desenvolvimento de câncer. Por exemplo, cerca de 5% dos cânceres de mama são atribuídos à herança de uma forma particular de *gene de suscetibilidade ao câncer de mama* [7]. Outras mutações herdadas foram descritas como aumentando o risco de uma pessoa a desenvolver câncer de cólon, rim, osso, pele e outras formas específicas de câncer. Mas, a essas condições herdadas, acredita-se, são atribuídos 10%, ou menos, dos casos de câncer [8].

As mutações associadas ao câncer, sejam somáticas ou de células germinativas, de mutação pontual ou de grandes inserções/remoções, alteram proteínas chaves em suas funções no biossistema humano. Uma grande variedade de mutações parece estar envolvida. Até mutações em regiões não codificantes, como regiões promotoras e reguladoras (indutoras e repressoras), podem resultar em supra- ou sub-expressão de proteínas necessárias à normalidade [9]. Coletivamente, estas mutações contribuem para a alteração do genoma de normal ao canceroso.

O ciclo celular é um processo crítico pelo qual a célula passa até copiar-se exatamente. A maioria dos cânceres apresentam mutações no sinal que regula os ciclos celulares de crescimento e divisão. A divisão celular normal é necessária para a geração de novas células durante o desenvolvimento, e para a substituição de células velhas, quando da morte das mesmas. A maioria das células permanece na intérfase, o período entre as divisões celulares, por pelo menos 90% do ciclo celular. A primeira parte da intérfase é chamada de $G1$ (indicando o primeiro intervalo, ou *gap*), seguida da fase S (de síntese do DNA), e então $G2$ (segundo intervalo). Durante $G1$, há rápidos crescimento e atividade metabólica, incluindo a síntese de RNA e proteínas. O crescimento celular continua durante a fase S, e o DNA é replicado. Na $G2$, há uma continuidade do crescimento e preparo para a divisão celular. A divisão celular (mitose) é denominada de fase M. Células que não se dividem por longos períodos não replicam seu DNA e são consideradas como estando na fase $G0$. Em células normais, os *genes supressores de tumor* atuam "freando" os sinais durante $G1$, parando ou diminuindo o ciclo celular antes da fase S. Os *genes de reparo do DNA* estão ativos durante todo o ciclo celular, particularmente durante $G2$, após a replicação do DNA, e antes do preparo do cromossomo para a mitose.

Os genes supressores de tumor [10] são uma família de genes, que instruem a célula a produzir proteínas que atenuam o crescimento e divisão celulares. Alguns o fazem pela codificação de fatores de transcrição para outros genes necessários ao crescimento lento. Por exemplo, o produto do gene supressor $TP53$ é chamado de proteína $p53$ [11].

Esta proteína liga-se diretamente ao DNA e leva à expressão de genes que inibem o crescimento ou disparam o mecanismo de morte celular programada (apoptose). Outros genes supressores de tumor codificam para proteínas que auxiliam o controle do ciclo celular. Ambas as cópias de um gene supressor de tumor devem estar ausentes, ou mutadas, para que ocorra o câncer. Uma pessoa portadora da mutação herdada em um gene supressor de tumor tem apenas uma cópia funcional do gene em todas as células. Para essa pessoa, a perda ou mutação da segunda cópia do gene em uma dessas células pode levar ao câncer.

Algumas mutações associadas ao câncer parecem envolver a falha de um ou muitos dos sistemas de reparo celular. Um exemplo de tal erro envolve o reparo do erro de pareamento do DNA [10, 12]. Após o DNA copiar-se a si mesmo, os produtos dos *genes de reparo do DNA* atuam como leitores de prova, identificando e corrigindo erros de pareamento. Se a perda ou mutação ocorre em genes de reparo do DNA, mutações esporádicas ocorrerão com maior freqüência, e se acumularão [13]. Outras mutações associadas ao câncer envolvem mutações nos *proto-oncogenes*, genes normalmente envolvidos na regulação do crescimento celular controlado. Estes genes codificam para proteínas que funcionam como fatores de crescimento, receptores de fatores de crescimento, moléculas de sinalização celular e fatores de transcrição nuclear—proteínas, que ligam-se aos genes para iniciar a transcrição. Quando um proto-oncogene sofre mutação, ou está supra-regulado, é chamado de *oncogene* e resulta no descontrole dos crescimento e transformação celulares [14, 15, 10]. No nível celular, apenas uma mutação, em um único alelo, é necessária para disparar o papel oncogênico no desenvolvimento do câncer.

Enquanto os primeiros suspeitos das mutações associadas ao câncer são os oncogenes, genes supressores de tumor e genes de reparo do DNA, há ainda mutações em genes que ativam e desativam carcinógenos, governam a diferenciação celular e outras que permitem o câncer invadir e gerar metastase em partes distantes do corpo [16].

O câncer é capaz de se espalhar pelo corpo por dois mecanismos: invasão e metastase [16]. Invasão refere-se à migração direta e penetração por células cancerosas nos tecidos circunvizinhos, enquanto a metastase refere-se à habilidade das células cancerosas de penetrar nos vasos linfáticos e sanguíneos, circularem por essas vias, e então invadirem tecidos em outras partes do corpo. Metástases são a maior causa de morte por doenças malignas [16]. Dependendo da presença ou ausência da capacidade de se espalhar por invasão ou metastase, os tumores são classificados como benignos ou malignos. Tumores benignos são tumores que não se espalham por invasão ou metastase; portanto, crescem

apenas localmente. Tumores malignos são tumores capazes de se espalharem por invasão ou metastase. Por definição o termo câncer é apenas aplicado a tumores malignos, ou malignidades.

Em adição a todas alterações moleculares que ocorrem em uma célula cancerosa, o ambiente entorno do tumor também sofre grandes alterações [15, 9]. A célula cancerosa perde receptores, que normalmente responderiam à sinalização de células vizinhas para uma parada do crescimento. Ao invés, os tumores auto-amplificam (regulação positiva) seu suprimento de sinais de crescimento. Eles ainda inundam a vizinhança com outros sinais, denominados citocinas, e enzimas, denominadas proteases. Esta ação destrói tanto a membrana basal quanto a matriz intersticial, que se interpõem entre o tumor e sua via metatástica; freqüentemente um vaso sanguíneo ou um duto do sistema linfático.

O câncer se desenvolve em diversas fases, dependendo do tipo de tecido afetado. De forma típica, estas fases são: *displasia*; *câncer in situ*, quando o exame microscópico revela células com características de câncer; *câncer invasivo localizado*, quando o crescimento celular anormal atinge áreas adjacentes do tecido de origem; *envolvimento do linfonodo* (órgão linfóide local ou secundário); e, *metástase distante*.

O crescimento do câncer pode envolver os vasos sanguíneos e causar sangramento, que será aparente se o câncer atingir parte de um órgão com contato direto ou indireto com o exterior do organismo. Por exemplo, pode haver sangue no catarro em um câncer de pulmão, sangue nas fezes em um câncer de cólon, ou sangue na urina em um câncer de bexiga. O crescimento do câncer pode também causar alterações funcionais. O câncer no sistema nervoso central, por exemplo, pode dar origem a sinais e sintomas neurológicos. No câncer avançado, um dos sintomas mais severos é a dor induzida pela compressão dos nervos.

Com mais de 10 milhões de novos casos por ano, o câncer vem se tornando uma das doenças mais incidentes do mundo [1]. As causas e tipos de malignidades humanas variam em regiões geográficas e tipos populacionais distintos, mas, na maioria dos países, dificilmente encontramos uma família sem uma vítima de câncer.

O *World Cancer Report* [1]—documento da OMS [1], que, entre outros, descreve a freqüência, tendências de incidência e mortalidade e causas conhecidas do câncer humano em diversos países—afirma que a razão de câncer pode aumentar em 50%, atingindo 15

[1] Organização Mundial de Saúde

milhões de novos casos, entre os anos 2000 e 2020. Há claras evidências de que o estilo de vida saudável e ações de saúde governamentais e de profissionais de saúde podem impedir esta tendência, prevenindo até $1/3$ dos casos de câncer no mundo. Exemplos de áreas onde ações podem fazer a diferença no retardo do aumento da razão de câncer e prevenção são: redução do consumo de tabaco (fumo), adoção de estilo de vida e dieta saudáveis, e detecção precoce através de rastreamento, principalmente, nos cânceres de colo de útero e mama.

No ano de 2000, tumores malignos foram responsáveis por 12% das quase 56 milhões de mortes no mundo [1]. Em muitos países, mais de $1/4$ das mortes foram atribuídas ao câncer. Em 2000, $5,3$ milhões de homens e $4,7$ milhões de mulheres desenvolveram tumores malignos e, conjuntamente, $6,2$ milhões morreram em conseqüência da doença. Neste mesmo ano, o câncer apareceu como primeiro problema de saúde pública em países em desenvolvimento, alcançando os efeitos encontrados nos países desenvolvidos/industrializados [1].

Em países desenvolvidos, a probabilidade de ser diagnosticado com câncer é maior que duas vezes a de países em desenvolvimento [1]. Naqueles países, cerca de 50% dos pacientes com câncer morrem em conseqüência da doença, enquanto que em países em desenvolvimento 80% das vítimas de câncer são diagnosticadas em estádios tumorais terminais e incuráveis, o que enfatiza a necessidade de melhores programas de detecção.

As malignidades mais fatais diferem das três formas mais prevalentes (tab. 1.1), com o câncer de pulmão responsável por $17,8\%$ de todas as mortes por câncer, estômago, $10,4\%$ e fígado, $8,8\%$.

No Brasil [2], as estimativas para o ano de 2006 apontavam para mais de 472 mil novos casos de câncer. A incidência em homens e mulheres, segundo a localização primária, pode ser vista na tabela 1.2.

A compreensão da genômica do câncer permanece um grande desafio. Biotecnologias como os microarranjos de DNA, que são capazes de determinar os perfis de expressão de centenas de genes simultaneamente, apesar de não substituírem, ao menos no futuro próximo, o exame histopatológico no diagnóstico primário do câncer, podem, em algumas situações, fornecer informações diagnósticas mais detalhadas [17, 18], classificar ou diferenciar entre tipos de malignidades morfologicamente similares [19], auxiliar na identificação de cânceres metastáticos de origem primária desconhecida e monitorar

Tabela 1.1: Prevalência mundial aproximada do número de casos novos por câncer segundo localização primária [1].

Localização Primária **Neoplasia Maligna**	**Estimativa** **Novos Casos/ano**
Pulmão	$1,2$ milhão
Mama	1 milhão
Cólon e Reto	940 mil
Estômago	870 mil
Fígado	560 mil
Colo de Útero	470 mil
Esôfago	410 mil
Cabeça e Pescoço	390 mil
Bexiga	330 mil
Linfoma Maligno não-Hodgkin	290 mil
Leucemia	250 mil
Próstata e Testículo	250 mil
Pancreático	216 mil
Ovariano	190 mil
Renal	190 mil
Endometrial	188 mil
Sistema Nervoso	175 mil
Melanoma	133 mil
Tiróide	123 mil
Faringe	65 mil
Doença de Hodgkin	62 mil
Subtotal	7,4 milhões
Outras Localizações	-
Todas as Neoplasias	> 10 milhões

Tabela 1.2: Estimativa nacional para o ano 2006 de número de casos novos por câncer, em homens e mulheres, segundo localização primária [2].

Localização Primária	**Estimativa dos Casos Novos**		
Neoplasia Maligna	**Masculino**	**Feminino**	**Total**
Mama Feminina	-	48.930	48.930
Traquéia, Brônquio e Pulmão	17.850	9.320	27.170
Estômago	14.970	8.230	23.200
Colo do Útero	-	19.260	19.260
Próstata	47.280	-	47.280
Cólon e Reto	11.390	13.970	25.360
Esôfago	7.970	2.610	10.580
Leucemias	5.330	4.220	9.550
Cavidade Oral	10.060	3.410	13.470
Pele Melanoma	2.710	3.050	5.760
Outras Localizações	61.530	63.320	124.850
Subtotal	179.090	176.320	355.410
Pele não Melanoma	55.480	61.160	116.640
Todas as Neoplasias	234.570	237.480	472.050

onde, e como, o genoma do câncer foi atingido durante terapias moleculares [20, 21]. Os microarranjos de DNA devem também permitir a procura sistemática para genes órgão- ou tecido-específicos.

O conhecimento adquirido da mineração e divisão dos dados de microarranjos são ferramentas importantes no auxílio à integração das informações sobre alterações fenotípicas e genotípicas, que ocorrem em um biossistema durante o progresso do câncer, o que, espera-se, seja convertido em intervenções melhores e mais precoces em pacientes com câncer. Talvez, a maior vantagem dos microarranjos de DNA sobre as abordagens tradicionais seja a possibilidade de aumentar os valores de predição positivos dos testes diagnósticos, conjuntamente à diminuição da proporção de resultados falso-positivos e falso-negativos, ou seja, o aumento da especificidade e sensibilidade do diagnóstico [22].

Dados oriundos de microarranjos de DNA são característicamente esparsos, com pequeno número de amostras em um grande espaço de atributos (perfis genicos), e contém ruídos *técnicos* e *biológicos* [23]. Os ruídos técnicos podem ser introduzidos em diferentes estágios, como a produção do arranjo, preparo das amostras, hibridização entre o cDNA tecidual e as sondas, extração do sinal e análise dos resultados da hibridação. Os ruídos biológicos advém da não-uniformidade genética das amostras comparadas, de impurezas, ou ainda, de classificações clínicas equivocadas.

Quando a análise de microarranjos envolve a classificação de padrões de expressão amostrais, como é comum na pesquisa do câncer, por exemplo, na discriminação de duas formas tumorais de difícil diagnóstico, ou mesmo entre células saudáveis e malignas, um desafio é a seleção de uma assinatura de perfis gênicos preditiva. A maioria dos genes é irrelevante à discriminação entre as diferentes amostras teciduais e introduzem ruído no processo de classificação, o que potencialmente diminui a contribuição daqueles preditivos.

Métodos desenvolvidos para a seleção de atributos são divididos em dois grandes grupos [24]: (1) métodos de filtragem; e, (2) métodos envelopados. Em microarranjos de DNA esses métodos são conhecidos, respectivamente, como abordagens de ordenação individual de genes e abordagens de ordenação de subconjuntos de genes. Na ordenação individual de genes destacam-se a correlação GS original [17], modificada [25] e híbrida [26]; a desordem [27]; a ordenação baseada na verossimilhança [28]; e, o TNoM [29]. Nas abordagens de ordenação de subconjuntos de genes o problema é reduzido a uma otimização onde o objetivo é encontrar o classificador, que com o menor

número de perfis gênicos classificará corretamente o maior número de amostras possível. Mesmo para um único classificador, a seleção gênica ótima pode apenas ser garantida pela busca exaustiva entre todos os possíveis subconjuntos gênicos. Dada a impossibilidade dessa busca devida à explosão combinatória de subconjuntos, alternativas como o emprego de heurísticas de eliminação seqüencial de genes ou algoritmos genéticos (GAs) foram relatadas.

Entre os métodos de eliminação retrógrada seqüencial destacam-se a eliminação recursiva de atributos, ou RFE [30], e sua versão baseada em uma medida de entropia, intitulada E-RFE [31, 32], ambas associadas a classificadores de Máquinas de Vetores de Suporte (SVMs). A classificação de perfis amostrais pode ser realizada com diferentes máquinas de aprendizado. As SVMs [33, 34] são classificadores binários que determinam hiperplanos coplanares de separação de margem ótima entre pontos próximos de cada classe (vetores de suporte), mesmo em dados esparsos, tendo sido empregadas com sucesso na análise de microarranjos de DNA [35, 36, 37, 38, 39, 30, 40].

Os GAs [41, 42] são abordagens que mantém uma população de soluções, possibilitando uma busca paralela e eficiente em espaços de atributos de grande dimensão, aparecendo como uma alternativa natural ao problema de otimização da seleção de subconjuntos de genes. Ainda que a otimização seja um problema de objetivos múltiplos, vários trabalhos optaram pela combinação desses em uma única função objetivo, tanto associados a classificadores SVM [43, 44, 45, 46, 40, 47], quanto a outros classificadores, como os de vizinhos próximos (kNN), Naïve Bayes (NB) e *weighted voting method* [48, 49, 50, 51]. Outros trabalhos empregaram algoritmos genéticos de objetivos múltiplos [52, 53, 54, 55].

Recentemente, algumas propostas apresentaram resultados promissores no aumento da eficiência ou diminuição do custo computacional na seleção de subconjuntos de genes por métodos evolutivos [43, 44, 40, 47]. Em FRÖHLICH *et al.* [43] e SOUZA e CARVALHO [40], algoritmos genéticos de objetivo único associados à otimização evolutiva do parâmetro de regularização das SVMs e estimativa do erro de generalização (validação cruzada) aproximado [56] foram empregadas. Em PENG *et al.* [44], algoritmos genéticos de objetivo único associados a SVMs não-lineares reduziram o número inicial de genes a subconjuntos preditivos de tamanho entre 36 e 40 e, empregaram a eliminação recursiva como pós-processamento desses subconjuntos. Em HUERTA *et al.* [47] foi empregada uma pré-filtragem de genes preditivos por métodos *fuzzy*, seguida de um algoritmo genético de objetivo único associado a SVMs não-lineares, onde genes predi-

tivos, definidos por pontos de corte na acurácia de treinamento, foram arquivados para uma segunda etapa evolutiva. Nessa tese, algumas propostas de aumento na eficiência ou diminuição do custo computacional na seleção de subconjuntos de genes por métodos evolutivos previamente relatadas serão combinadas em um novo algoritmo de seleção de perfis gênicos em dados de microarranjos de DNA.

1.1 Objetivos

Nessa tese abordaremos o problema de identificação de subconjuntos de genes preditivos para a classificação binária em malignidades humanas. O objetivo desse trabalho é investigar um novo algoritmo de seleção de perfis gênicos em dados de microarranjos de DNA utilizando Máquinas de Vetores de Suporte (SVMs), Algoritmos Genéticos de Objetivos Múltiplos (MOGA) e Eliminação Recursiva de Atributos (RFE), intitulado RFE–SVM–MOGA. São propostas soluções específicas, como a inicialização semi-aleatória dos cromossomos do MOGA; otimização evolutiva da constante de regularização C das SVM; controle da diversidade genotípica populacional durante o processo evolutivo do MOGA; parada precoce do MOGA por critérios explícitos; utilização racional de perfis gênicos selecionados em diferentes simulações do SVM–MOGA pela aplicação do princípio da superioridade; e, eliminação recursiva (RFE) dos atributos selecionados pelo SVM–MOGA.

Para atingir o objetivo, o método proposto foi aplicado em três estudos com conjuntos de dados de microarranjos de DNA publicamente disponíveis, correspondendo aos estudos de classificação binária em Leucemia Aguda [17], Linfoma difuso de grandes células B (DLBCL) [19] e Câncer de Cólon [57].

Contribuições dessa tese incluem progressos na identificação de perfis gênicos preditivos, que é um desafio na análise de padrões amostrais em dados de microarranjos de DNA; melhor compreensão de algoritmos de seleção em problemas esparsos; e, recomendação de escolhas apropriadas de algoritmos de seleção de perfis gênicos para problemas específicos.

1.2 Estrutura do Trabalho

No capítulo 2 são apresentados os fundamentos teóricos dos microarranjos de DNA e uma revisão bibliográfica da aplicação dessa técnica de biologia molecular na pesquisa do câncer humano. Na descrição da tecnologia (seção 2.1) são apresentadas uma comparação dos microarranjos de DNA com métodos experimentais clássicos como o *Northern Blotting*; as plataformas mais empregadas, acompanhadas de uma descrição sucinta dos passos envolvidos em suas fabricações; e, a matriz de dados oriunda de experimentos de microarranjos de DNA. Na seção 2.2, é apresentada uma extensa revisão da literatura do emprego de microarranjos de DNA na pesquisa do câncer.

No capítulo 3 são apresentados os fundamentos teóricos dos algoritmos genéticos de um (GA) ou múltiplos objetivos (MOGA); uma extensa revisão bibliográfica da seleção de atributos em microarranjos de DNA por métodos evolutivos ou não-evolutivos. Na apresentação dos GA (seção 3.1), é dada uma visão geral do processo de busca evolutiva e, é descrito o GA canônico, assim como seus principais operadores binários (seleção, mutação, recombinação e reinserção). Na seção 3.2 são apresentados a otimização de objetivos múltiplos e os algoritmos genéticos desenvolvidos para esse propósito, assim como a ordenação pelo princípio da superioridade das soluções e métodos de manutenção da diversidade no processo evolutivo. A seguir (seção 3.3), fazemos uma extensa revisão dos métodos de filtragem e envelopados empregados na seleção de perfis gênicos. Ainda nesse capítulo (seção 3.4), destacamos o emprego de GAs na seleção gênica em dados de microarranjos de DNA apresentando uma revisão dos métodos e dos resultados relatados na literatura.

No capítulo 4, apresentamos a classificação de perfis de expressão amostral por Máquinas de Vetores de Suporte (SVMs). Primeiramente, são apresentadas as bases teóricas das SVMs (seção 4.1), onde o treinamento de sua apresentação mais simples (SVM lineares), assim como de sua extensão (C–SVM) que possibilita a obtenção de margens de classificação *soft*. São apresentadas linhas gerais da Teoria Vapnik-Chervonenkis importante para a compreensão da capacidade de generalização dessas máquinas de aprendizado mesmo em amostras esparsas. A seguir, destacamos os principais trabalhos com emprego de SVMs na classificação de perfis de expressão amostral em dados de microarranjos de DNA.

No capítulo 5, é apresentado o algoritmo proposto nessa tese, denominado RFE–SVM–MOGA, que associa um MOGA com classificadores C–SVMs especialmente desenhado para a seleção de atributos em microarranjos de DNA, de treinamento rápido, seguido da Eliminação Recursiva de Atributos (RFE). No capítulo 6, avaliamos o método proposto em três estudos envolvendo dados públicos de microarranjos de DNA em malignidades humanas (seções 6.1, 6.2 e 6.3). Os resultados obtidos são comparados com outros próprios assim como relatados na literatura e discutidos no capítulo 7.

Ao final (capítulo 8), apresentamos um resumo da tese, recomendações finais e algumas propostas de trabalhos futuros.

Capítulo 2

Microarranjos de DNA na Pesquisa do Câncer

2.1 Microarranjos de DNA

A base fundamental dos microarranjos de DNA é o processo de *hibridização* competitiva ou simultânea. Duas fitas de DNA hibridizam, se são complementares entre si, ou seja, se há pareamento entre suas bases. Uma, ou ambas fitas de DNA, podem ser substituídas por RNA e continuará havendo hibridização, enquanto houver complementaridade.

Há décadas, a hibridização é usada na biologia molecular como base de técnicas tais como o *Southern Blotting* [58] e o *Northern Blotting* [59]. No *Southern Blotting*, uma pequena seqüencia de DNA, um *oligonucleotídeo*, é utilizado para hibridizar com fragmentos complementares de DNA, previamente separados por tamanho em um gel de eletroforese. Se o oligonucleotídeo está marcado com radiação (radioisótopo), pode-se observar a hibridização em filme fotográfico (sensível à radiação). No *Northern Blotting* um oligonucleotídeo marcado por radiação é utilizado na hibridização com RNA mensageiro (mRNA), também previamente submetido à corrida em gel. A quantidade de radiação capturada em filme fotográfico é dependente, até certo ponto, da quantidade de sonda marcada com radiação, que novamente, é dependente da quantidade de mRNA (ensaio semi-quantitativo).

Os arranjos de DNA são versões seriadas maciças dos *Northern* e *Southern Blotting* (fig. 2.1). Ao invés de se distribuir sondas de RNA ou DNA em um gel contendo

amostras de RNA ou DNA, as sondas de oligonucleotídeos são imobilizadas em uma superfície. Estas podem ser depositadas em distâncias micrométricas, tornando possível a deposição de centenas de sondas diferentes de oligonucleotídeos em uma única e pequena superfície de $1\,\mathrm{cm}^2$. Ao invés de utilizar radioatividade, a amostra é geralmente marcada com corantes fluorescentes [1]. Esses corantes emitem fótons quando excitados pela luz—correspondente ao material hibridizado—, o que pode ser detectado por um *scanner* confocal a laser.

Onde antes era possível realizar um par de *Northern Blots*, ou um par de *Southern Blots*, agora, em um único dia, realizam-se, com os arranjos de DNA, dezenas de milhares de hibridizações. Disto, resultou uma revolução na biologia molecular, e por conseqüência, na Medicina. Ao invés de se estudar poucos genes e mRNAs por vez, cientistas estão agora estudando muitos genes e mRNAs simultaneamente. Os arranjos de DNA são rotineiramente usados no estudo de todos os mRNAs conhecidos de um organismo [60]. Abriu-se então, a possibilidade de uma visão sistemática, completamente nova, da reação de células a um certo estímulo [61], assim como, do estudo das doenças humanas a partir da observação do efeito dessas na expressão global dos genes celulares [62].

Há diversas plataformas descritas para estudos de microarranjos de DNA [63, 64]. Os tipos de arranjos mais usados atualmente são os arranjos de cDNA e os arranjos sintetizados *in situ*, ou *genechips*. Os arranjos de cDNA foram primeiro descritos em um estudo de expressão envolvendo 45 genes de *Arabidosis* spp., desenvolvido por SCHENA *et al.* [65], onde a expressão diferencial foi medida por meio de fluorescência de hibridização de duas cores. No ano seguinte, DERISI *et al.* [66] aplicaram essa tecnologia no estudo de mais de 1000 genes no melanoma. Já os arranjos de oligonucleotídeos sintetizados *in situ* foram primeiro descritos por LOCKHART *et al.* [67] e estão sendo agora disponibilizados comercialmente pela *Affymetrix* [64, 68, 69].

Os microarranjos de cDNA (fig. 2.2) utilizam um robô para mover pequenos volumes de sonda em solução—DNA complementar (cDNA), oligonucleotídeos pré-sintetizados, ou produtos amplificados pela técnica da reação em cadeia da polimerase (PCR)—de um placa microtitulada para a superfície de uma lâmina de vidro. Cada sonda é complementar a um único gene. As sondas podem ser fixadas de diversas formas, sendo a mais clássica a união não específica à lâminas cobertas de poli-lisina. As etapas envolvidas na fabricação das lâminas podem ser sintetizadas como [70]: (1) Cobertura das

[1] Geralmente corantes de cianina ($Cy3$ e $Cy5$).

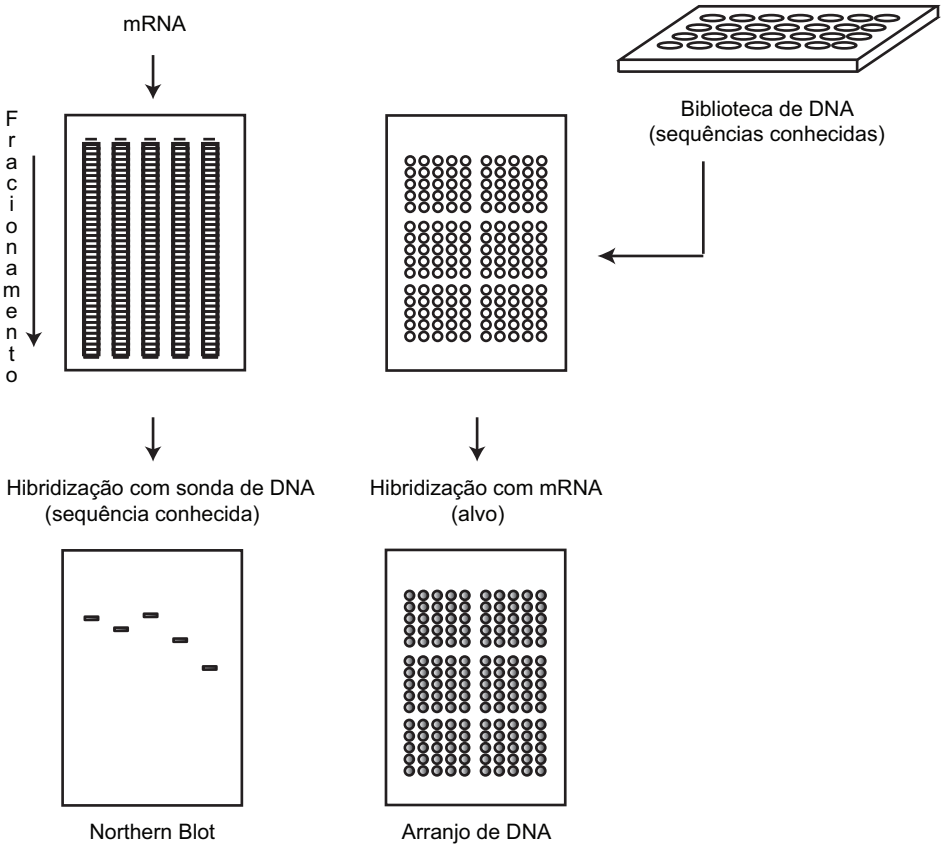

Figura 2.1: Comparação entre as tecnologias dos *Northern Blottings* e dos Microarranjos de DNA. No *Northern Blotting*, o mRNA extraído de uma célula é submetido a uma corrente elétrica, em um gel de agarose, ou poliacrilamida, onde é fracionado de acordo com seu peso molecular (eletroforese), e transferido para um filtro de papel (alvo), onde irá hibridizar com seqüencias de DNA conhecidas, obtidas de uma biblioteca genômica, marcadas usualmente com radioisótopos (sonda). Nos microarranjos de DNA, o mRNA extraído de uma célula é marcado com fluorescência (sonda), e hibridado com seqüencias de DNA conhecidas (alvo), obtidas de uma biblioteca genômica, arranjadas em uma lâmina de vidro (arranjos de cDNA), ou construídas em um *biochip* (arranjos de oligonucleotídeos). Além dessas diferenças, no *Northern Blotting* há a possibilidade de se testar um número de mRNA no máximo igual ao número de raias da cuba de eletroforese, menos um—ao menos uma raia é usada para o marcador de peso molecular, enquanto, nos microarranjos de DNA, esse número é de dezenas de milhares.

lâminas com poli-lisina; (2) Preparo das sondas em placas microtituladas; (3) Utilização do robô para arranjar as sondas em lâminas de vidro; (4) Bloqueio das aminas remanescentes da poli-lisina com anidrato sucínico; e, (5) Desnaturação do DNA (se de dupla-fita) por calor.

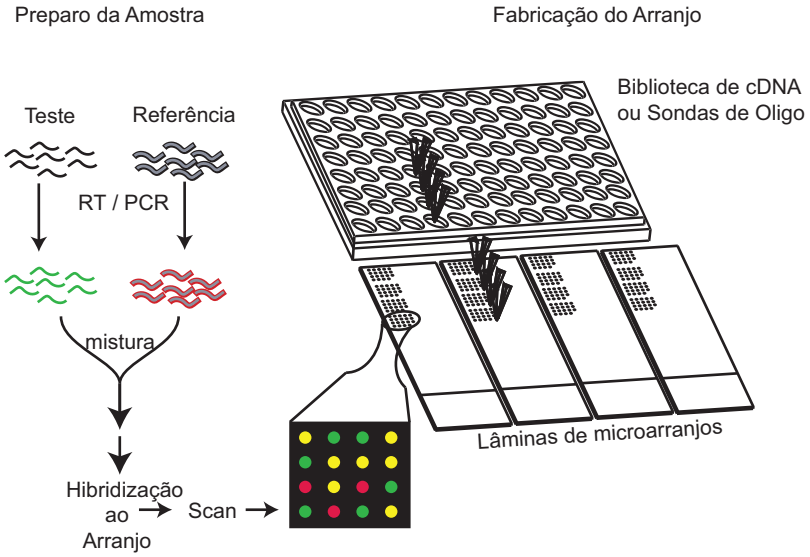

Figura 2.2: A tecnologia dos arranjos de cDNA. Um robô (*arrayer*) é usado na transferência das sondas em solução contidas em placa microtitulada, para uma lâmina de vidro. O mRNA extraído das células investigadas são convertidos a cDNA (transcrição reversa) e marcados com fluorescência. Usualmente (desenho de referência) a amostra de interesse é marcada em vermelho, enquanto a amostra de referência é marcada em verde. Após serem misturadas, estas são hibridizadas às sondas na lâmina de vidro. Após lavagem do material não-hibridizado, a lâmina é escaneada com um microscópio confocal a laser, e a imagem é computacionalmente analisada.

Da mesma forma, os passos envolvidos na preparação da amostra (técnicas de marcação indireta) e hibridização ao arranjo podem ser resumidas como [70] (fig. 2.2): (1) Extração do RNA total das células; (2) Isolamento do mRNA pelo método da cauda de poli–A (opcional); (3) Conversão do mRNA em cDNA na presença de aminoalil–dUTP (AA–dUTP); (4) Hibridização do mRNA marcado às lâminas de vidro; (5) Lavagem do material não hibridizado; e, (6) Varredura da lâmina e análise da imagem.

Já a *Affymetrix* usa equipamento similar àqueles utilizados na fabricação de *chips* de silício, utilizados em computadores pessoais, o que permite a produção em série de lâminas a um preço razoável. Onde os *chips* de computadores são feitos pela criação

de máscaras, que controlam o processo de fotolitografia para remoção ou depósito de materiais na superfície do *chip*, a *Affymetrix* usa máscaras para controlar a síntese de oligonucleotídeos na superfície da lâmina. O método da fosforamidita, padrão de síntese de oligonucleotídeos, foi modificado para permitir controle por luz de cada passo [71, 72]. As máscaras controlam a síntese de várias centenas de milhares de quadrados (células do arranjo), cada um contendo muitas cópias—geralmente mais de 40—do mesmo oligonucleotídeo [73, 74].

Diferente dos arranjos de cDNA, onde utilizam-se oligonucleotídeos, ou cDNAs, relativamente longos (50–70 bases de comprimento), nos arranjos da *Affymetrix* utilizam-se sondas de oligonucleotídeos curtas, com 25 nucleotídeos de comprimento ($25\,\mathrm{mer}$), normalmente selecionadas a partir da região $3'$ do RNA alvo—região suficientemente representativa de cada gene que, presume-se, têm o menor grau de similaridade com outros genes. Geralmente, de 16–20 oligonucleotídeos são escolhidos como PM (de *perfect matches*), ou perfeitamente pareados (i.e., com perfeita complementaridade com o mRNA daquele gene), enquanto 16–20 oligonucleotídeos MM (de *mismatches*) possuem uma única base sem pareamento exato, localizada no meio do oligonucleotídeo (fig. 2.3). A *Affymetrix* afirma que estes últimos são capazes de detectar hibridização não-específica e hibridização de fundo (ruído), que são importantes na quantificação de mRNAs de baixa expressão.

Generalizando, como mostrado na figura 2.4, os passos na construção de um arranjo de oligonucleotídeos sintetizados *in situ* são [73, 74]: (1) Extração do RNA total da célula (normalmente se utiliza TRIzol ou *kits RNeasy*); (2) Conversão do mRNA em cDNA, usando a transcriptase reversa e o *primer poly–T*; (3) Amplificação do cDNA resultante, usando a polimerase de RNA T7 na presença de biotina–UTP e biotina–CTP, para que cada cópia de cDNA produza de 50–100 cópias de cRNA (RNA complementar) marcados pela biotina; (4) Incubação do cRNA a 94^oC, em solução fragmentadora, gerando fragmentos de cRNA de 35–200 nucleotídeos de comprimento; (5) Hibridização ao *genechip* e lavagem do material não hibridizado; (6) Escaneamento do *genechip* com *scanner* confocal a laser; (7) Amplificação do sinal do *genechip* com IgG de cabra e anticorpo biotinilado; e, (8) Nova varredura do *genechip*.

A vantagem dos arranjos de cDNA, quando comparados aos arranjos de oligonucleotídeos da *Affymetrix*, é que se pode desenhar qualquer sonda para deposição no arranjo. As desvantagens seriam o custo, bem menor o da *Affymetrix*, e a uniformidade

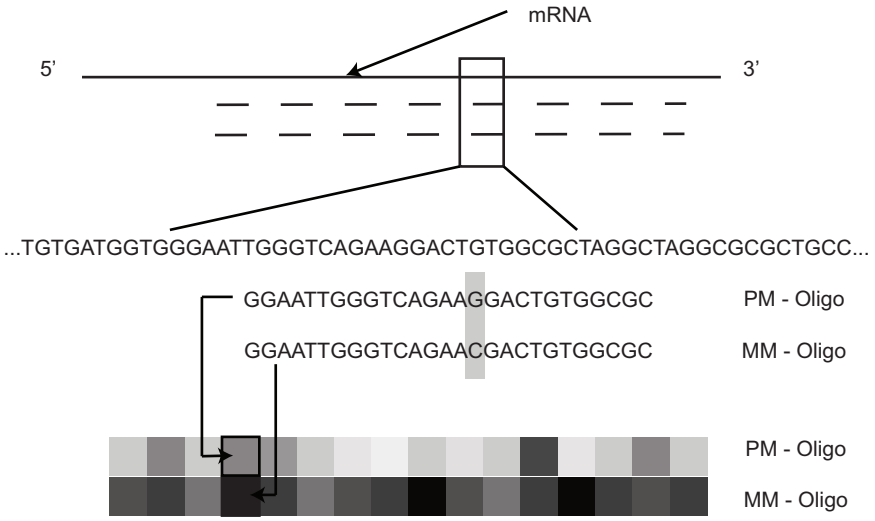

Figura 2.3: A tecnologia dos arranjos de oligonucleotídeos (*Affymetrix*). A presença de mRNA é detectada por uma série de pares de sondas, diferenciadas por um único nucleotídeo. A hibridização dos mRNAs fluorescentes com essas sondas no arranjo é detectada pelo escaneamento a laser da superfície do arranjo.

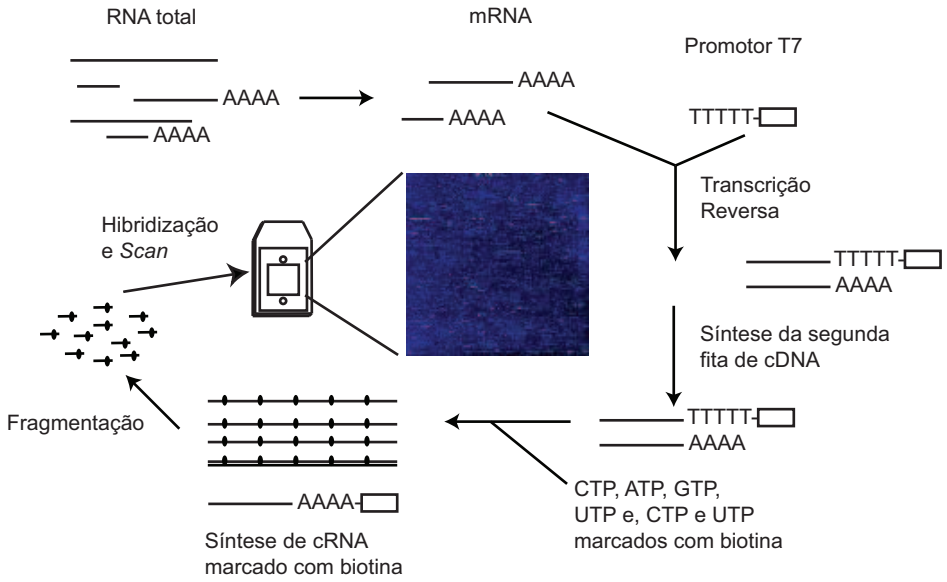

Figura 2.4: Preparo da amostra para os arranjos da *Affymetrix*. O mRNA é extraído da célula e convertido a cDNA. Há então passos de amplificação e marcação desse cDNA, anteriores à fragmentação e hibridização a oligonucleotídeos de $25\,\mathrm{mer}$ na superfície do arranjo. Após lavagem do material não-hibridado, o arranjo é escaneado em um *scanner* a laser confocal, e a imagem analisada em computador.

das lâminas. Do ponto de vista da análise dos dados, a maior diferença é que nos arranjos de cDNA, geralmente, a amostra e o controle são hibridizados na mesma lâmina, usando-se diferentes marcadores, enquanto nos *chips* da *Affymetrix* é possível usar apenas um marcador, sendo necessários dois *chips* para se comparar, por exemplo, tratamento e controle. Enquanto há evidências de que diferentes plataformas de microarranjos produzem perfis de expressão gênicos diferentes, não há indicação de que uma plataforma seja melhor que a outra [75, 76, 77].

Os dados de microarranjos são normalmente organizados como uma matriz $\mathbf{M}(n \times m)$, onde as linhas são n genes e as colunas m amostras biológicas que tiveram a expressão monitorada (fig. 2.5). Então, $\mathbf{M}_{ij}$ indica a expressão do gene i na amostra j. Seja $\mathbf{e}_i$, a i-ésima linha de $\mathbf{M}$, o padrão de expressão do gene i nas m amostras, pode-se referir a esse padrão/perfil como *padrão de expressão gênico*. Seja $\mathbf{s}_j$, a j-ésima coluna de $\mathbf{M}$, o nível de expressão dos n genes na amostra j, pode-se referir a esse padrão/perfil como *padrão de expressão amostral*.

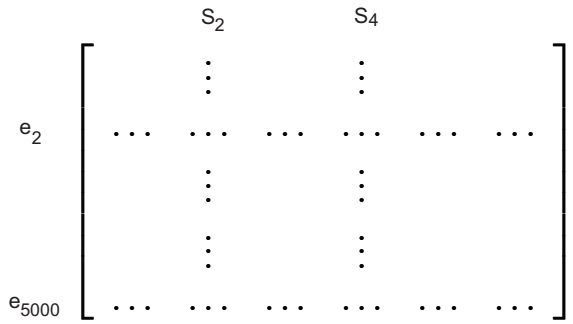

Figura 2.5: A Matriz de dados de microarranjos. São mostrados os padrões de expressão gênico dos genes 2 e 5000, e os padrões de expressão amostral das amostras 2 e 4.

A tecnologia de arranjos de cDNA produz dados onde a expressão e_{ij}, do gene i na amostra j, é calculada como a razão r_{ij}/g_i do nível real de expressão do gene i, r_{ij}, na amostra j, sobre o nível de expressão do gene i em uma amostra controle, ou ainda, outra amostra de interesse, g_i. Além disso, visando a normalização dos dados—aproximação de uma distribuição normal dos dados—é comum transformar-se esse valor pelo $\log$,

$$e_{ij} = \log \frac{r_{ij}}{g_i}. \tag{2.1}$$

Quando da visualização dos dados (fig. 2.6), o valor de e_{ij} é codificado (pseudo-colorido) por cores em uma mistura de vermelho e verde (vermelho quando $r_{ij} \gg g_i$, verde quando $r_{ij} \ll g_i$, e uma mistura (amarelo) neste intervalo).

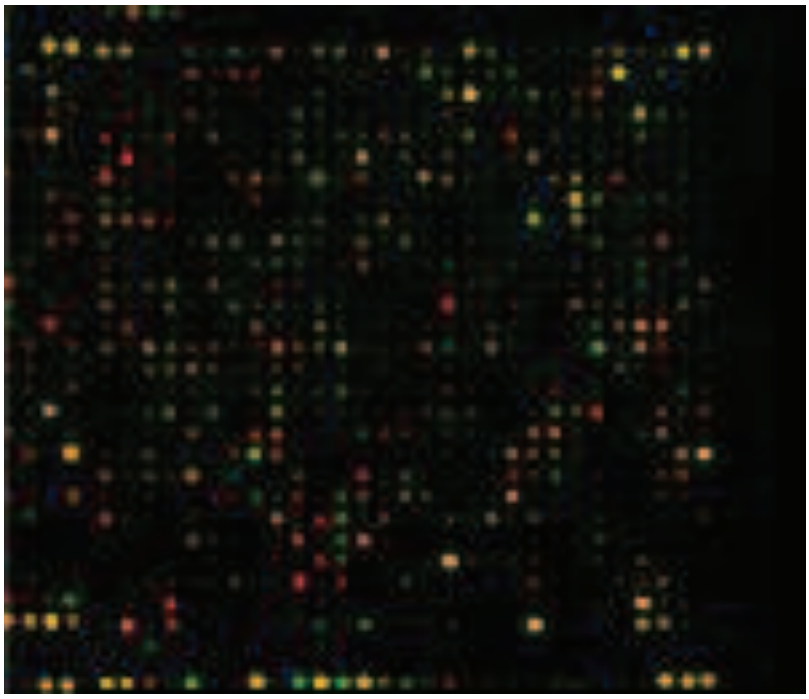

Figura 2.6: Composição de uma imagem pseudo-colorida de microarranjos de cDNA. As duas imagens de intensidade de 16-`bits`, escaneadas de uma lâmina de microarranjo de DNA, cada qual com comprimento de onda de emissão correspondente ao intervalo de excitação dos corantes de cianina usados ($Cy3$ e $Cy5$), compõem uma imagem RGB, onde os valores de intensidade de cada canal/lâmina é convertido em intensidades (valores no intervalo $[0; 65536]$) das matizes vermelho e verde. Como resultado, temos uma imagem, onde um *spot* (pontos da imagem) vermelho indica que o mRNA da amostra corada por $Cy3$ tem intensidade de expressão (correspondente ao número de transcritos) muito maior, que na amostra corada por $Cy5$; um *spot* verde indica, que o mRNA da amostra corada por $Cy5$ tem intensidade de expressão maior, que na amostra corada por $Cy3$; e, um *spot* amarelo indica que o mRNA de ambas amostras tem intensidade de expressão similar.

No caso dos arranjos da *Affymetrix*, os pares de sondas *perfect matches* (PM) e *mismatches* (MM) associados são analisados e usados na produção de diversas estatísticas relacionadas ao nível de expressão gênico. A medida usada com mais freqüência é a AVD (de *Average Difference*), que é derivada das diferenças entre os 16–20 pares

de sonda PM e os 16–20 pares de sonda MM dos n oligonucleotídeos utilizados na detecção de cada gene

$$e = AVD = \frac{\sum_{i=1}^{n} PM - MM}{n}, \tag{2.2}$$

O *software* da *Affymetrix* também gera uma chamada de presente/ausente para cada gene no *genechip*. A função dessa chamada é indicar quando um gene está ausente (A), presente (P), ou marginal (M). Entretanto, espera-se que essa ausência, ou presença, do gene seja revelada pelos dados, normalmente ignorando-se essa função.

2.2 Microarranjos de DNA na Pesquisa do Câncer

Um primeiro problema na genômica do câncer abordado por microarranjos de DNA, foi a procura de *marcadores de tumor* para fins diagnósticos (tab. 2.1). Estes, idealmente, devem se expressar especificamente em tumores ou em tecidos pré-malignos; mostrar pouca ou nenhuma expressão em tecidos normais ou benignos; serem produzidos, especificamente, por um órgão; e, serem medidos por ensaios simples, baratos, padronizados e reprodutíveis [78]. Além disso, os marcadores devem estar presentes em amostras biológicas obtidas, preferencialmente, por procedimentos não, ou pouco, invasivos. Os marcadores de tumor atualmente existentes carecem de especificidade para malignidades e, com exceção do antígeno específico da próstata (*Prostate Specific Antigen*; PSA), não são órgão específicos. Nenhum dos marcadores atualmente em uso, com exceção da gonadotrofina coriônica humana (*Human Chorionic Gonadotrophin*; HCG), estão alterados em todos os tumores de um tipo específico.

Em uma publicação pioneira, GOLUB *et al.* [17] analisaram a expressão de 7129 genes na medula óssea de 38 pacientes com leucemia aguda (27 com a forma linfoblástica aguda (ALL), e 11 com a forma mielóide aguda (AML)). Foram selecionados 50 genes que apresentaram maior diferenciação de níveis de expressão entre as células AML e ALL. Usando-se a assinatura genômica composta por esse subconjunto de genes, os investigadores foram capazes de identificar corretamente quais pacientes tinham AML e quais tinham ALL em um novo estudo de coorte com 34 pacientes. Apesar de a diferenciação entre esses dois tipos de leucemia não ser um problema de difícil diagnóstico, este estudo foi um dos primeiros a mostrar que o perfil de expressão gênico poderia ser usado na classificação de malignidades. Como indicador do impacto desse estudo, o conjunto

Tabela 2.1: Exemplos de trabalhos usando microarranjos de DNA na procura de *marcadores de tumor* para fins diagnósticos.

Malignidades	**Publicação**
Leucemia linfoblástica aguda (ALL) e Leucemia mielóide aguda (AML)	GOLUB *et al.* [17]
Tumores de células pequenas, redondas e azuis (SRBCTs)	KHAN *et al.* [18]
Oligodendroglioma anaplástico e glioblastoma	NUTT *et al.* [79]
Tumores metastático de origem diversa	SU *et al.* [80]
Tumores metastático de Pulmão, Cólon e Ovário	GIORDANO *et al.* [81]

de dados de GOLUB *et al.* [17] tornou-se um dos mais re-analisados conjuntos de dados de microarranjos de DNA [37, 82, 83, 84]. Os mesmos 6 genes mais discriminantes entre as formas ALL e AML foram relatadas por GRANT *et al.* [82] e CULHANE *et al.* [83]. Entretanto, apenas 13 dos 50 genes mais discriminantes entre AML e ALL identificados por CULHANE *et al.* [83], foram também encontrados por GOLUB *et al.* [17]. Também foi notado que houve problemas na classificação de uma única amostra de teste; foi sugerido que essa amostra estava erroneamente rotulada.

Diferente da situação com AML e ALL discutida acima, um grupo de tumores conhecido como *Tumores de células pequenas, redondas e azuis* (SRBCTs), são particularmente de difícil classificação. Esta coleção de malignidades pediátricas inclui o neuroblastoma, rabdomioblastoma, tumor de Ewing e linfoma de Burkitt, um tipo de linfoma não-Hodgkin. O diagnóstico preciso desses tumores é essencial tanto ao prognóstico quanto para a terapia tumor-dependente. Na tentativa de sub-classificar, KHAN *et al.* [18] mediram a expressão de mais de 6500 genes usando microarranjos de cDNA. Primeiro, o número de genes examinados foi reduzido a 96, usando-se rondas iterativas de Análise de Componentes Principais (PCA). Segundo, usando apenas os 96 genes melhor ordenados, os pesos de conexão de uma rede neural artificial foram otimizados para a classificação de 25 amostras de teste. Os perfis amostrais de 5 tumores não-SRBCT e 20 tumores SRBCT foram corretamente classificados, apesar de apenas 17 das 20 SRBCTs terem sido originalmente diagnosticadas com certeza.

Assim como na leucemia aguda e nos SRBCTs, entre os gliomas de alto-grau, o oligodendroglioma anaplástico apresenta um prognóstico mais favorável, que o glioblastoma. Além disso, apesar de os glioblastomas serem resistentes a maioria das terapias em uso, os oligodendrogliomas anaplásticos são, normalmente, quimiosensíveis. O diagnóstico preciso permite decidir qual a gerência mais adequada dos indivíduos. NUTT *et al.* [79] utilizaram microarranjos de DNA na procura de marcadores tumorais, que diferenciassem essas duas formas de glioma de alto-grau.

Outro problema diagnóstico freqüente no laboratório de patologia clínica é a identificação de cânceres metastáticos de sítios primários desconhecidos. A metástase de origem desconhecida é definida como uma metástase para o qual o sítio primário permanece desconhecido apesar da obtenção de história clínica, exame físico, radiografia pulmonar, análises rotineiras de urina e sangue, conjuntamente com exames microscópicos e histológicos. A metástase de origem desconhecida abrange 5–10% de todos os cânceres. O reconhecimento do sítio primário é essencial para a escolha da proposta terapêutica mais adequada. Na tentativa de abordar esse problema, SU *et al.* [80] usaram a informação obtida da análise de expressão gênica de 100 tumores primários como conjunto de treinamento. Outros 75 tumores primários e metastáticos foram usados como amostra de teste. A acurácia da predição de 97% e 95% foram obtidas, respectivamente, nos conjuntos de treinamento e teste. Com apenas 11 genes, os autores puderam predizer a origem anatômica de 91% e 83% dos tumores metastáticos na amostra de treino e teste, respectivamente. Usando uma abordagem similar, GIORDANO *et al.* [81] classificaram corretamente $152/154$ amostras de adenocarcinoma, de acordo com sua origem, derivadas de pulmão, cólon e ovário.

BLOOM *et al.* [85] empregaram redes neurais artificiais (ANNs) tanto em microarranjos de cDNA como de oligonucleotídeos. Na classificação de tumores capazes de discriminar entre 8 diferentes tipos de adenocarcinoma de histopatologia similar, os autores utilizaram um arranjo de cDNA de $32\,\mathrm{k}$ e obtiveram acurácia média de classificação de 83%. Para a avaliação das ANNs em arranjos de oligonucleotídeos, em 463 tumores representativos de 21 tipos tumorais, o treinamento foi efetuado com uma amostra aleatória contendo 343 tumores, e a acurácia de predição de origem patológica na amostra de teste (120 tumores) foi de 88%. Em uma plataforma contendo amostras tanto de microarranjos de cDNA quanto de oligonucleotídeos, 539 tumores foram divididos em amostras aleatórias de treinamento e teste, obtendo acurácia de teste de 85%. Resultado bastante

similar (acurácia de teste de 84%) foi obtida na predição do sítio de origem de 50 amostras metatástica de cérebro, pulmão e fígado.

Seguindo a classificação das malignidades por análise de expressão gênica, outras questões clínicas foram abordadas, incluindo a determinação de *fatores prognósticos* e *fatores preditivos* (tab. 2.2). Os *fatores prognósticos* são características clínicas, patológicas e bioquímicas dos pacientes com câncer e de suas formas tumorais, enquanto os *fatores preditivos* incluem a resposta ou resistência à terapias específicas. Muitos estudos usando microarranjos de DNA na determinação dos fatores prognósticos, foram realizados na presença de drogas adjuvantes—drogas usadas na prevenção do câncer em populações de risco, ou, mais usualmente, na prevenção de recidiva do mesmo—tornando difícil a separação entre os dois fatores.

Como a maioria dos cânceres, o *Linfoma difuso de grandes células B* (DLBCL) apresenta heterogeneidade clínica. Enquanto quase todos os pacientes com DLBCL apresentam doença agressiva, 35–40% podem ser curados pelo tratamento quimioterápico com o antibiótico anti-neoplásico antraciclina. ALIZADEH *et al.* [19] relacionaram o prognóstico da DLBCL com o padrão de expressão gênico do tumor. Usando microarranjos de DNA contendo aproximadamente 4000 clones de cDNA, conjuntamente com uma abordagem não-supervisionada de análise, foram identificadas duas formas de DLBCL, uma tendo o perfil de expressão gênico similar àquele do linfoma angiocêntrico de células B, enquanto outros tinham um padrão característico de células B periféricas ativadas *in vitro*. Pacientes com o último perfil apresentaram um prognóstico pior que àqueles com o primeiro perfil. Este estudo mostrou que o perfil de expressão gênico era capaz de subdividir uma única categoria diagnóstica de linfoma em dois subtipos distintos, cada um com prognóstico diferente.

Já SHIPP *et al.* [87] utilizaram arranjos de oligonucleotídeos e análise supervisionada na determinação do prognóstico de pacientes com DLBCL. Neste estudo, a expressão de aproximadamente 6800 genes foi analisada em 58 pacientes tratados com adriamicina, ciclofosfamida, vincristina e prednisolona. Desta análise, 13 genes foram encontrados como fatores de prognóstico, independente do Índice Prognóstico Internacional (IPI). O IPI é um índice prognóstico amplamente usado na determinação do prognóstico em pacientes com DLBCL, e é baseado em cinco características (idade, estádio clínico, níveis de desidrogenase lactica, índice de Karnofsky e número de sítios extranodais comprometidos).

Tabela 2.2: Exemplos de trabalhos usando microarranjos de DNA na determinação de *fatores preditivos* e *prognósticos* de formas tumorais.

Malignidade	**Publicação**
Linfoma difuso de grandes células B (DLBCL)	ALIZADEH *et al.* [19]
	RESENWALD *et al.* [86]
	SHIPP *et al.* [87]
	LOSSOS *et al.* [88]
Linfoma folicular	DAVE *et al.* [89]
Câncer de mama	VAN'T VEER *et al.* [90]
	WANG *et al.* [91]
	SORLIE *et al.* [92]
	AHR *et al.* [93]
	RAMASWAMY *et al.* [94]
	GLINSKY *et al.* [95]
	VAN DE VIJVER *et al.* [96]
Câncer de próstata	DHANASEKARAN *et al.* [97]
	SINGH *et al.* [98]
	GLINSKY *et al.* [99]
Câncer Renal	MOCH *et al.* [100]
Câncer de cólon e reto	WANG *et al.* [101]
	ESCHRICH *et al.* [102]
Carcinoma hepatocelular	IIZUKA *et al.* [103]
Carcinoma escamoso de cabeça e pescoço	ROEPMAN *et al.* [104]
Câncer de pulmão	BEER *et al.* [105]
Câncer esofágico	KIHARA *et al.* [106]
Blastoma medular	FERNANDEZ *et al.* [107]

Em outro estudo em DLBCL, RESENWALD *et al.* [86] mediram a expressão de mRNA em 240 pacientes, que tiveram tratamento subseqüente com quimioterapia baseada na antraciclina. Assim como no relato de SHIPP *et al.* [87], arranjos de oligonucleotídeos conjuntamente com análise supervisionada foram utilizados. Três grandes subgrupos foram identificados baseados em dados de perfil de expressão gênico. Um subgrupo exibiu padrão de expressão gênico característico de linfoma de grandes células B centroblásticas. O segundo subgrupo expressou características gênicas de linfoma de grandes células B ativadas e células B mitogenicamente estimuladas, enquanto o último grupo não apresentou características de expressão de nenhum desses grupos celulares. A translocação do gene *bcl-2* e a amplificação do gene *c-rel* foram encontradas apenas na sub-classe com características de células B centroblástica.

Os genes preditores identificados nos três estudos acima em DLBCL apresentaram pouca, ou nenhuma, superposição. Por exemplo, apenas 13 genes identificados por SHIPP *et al.* [87] estavam presentes no grupo gênico selecionado por ALIZADEH *et al.* [19], enquanto não houve qualquer superposição com os selecionados pelo modelo de ROSENWALD *et al.* [86].

O linfoma folicular é a segunda forma mais comum de linfoma não-Hodgkin, contribuindo com mais de 20% de todos os casos [108]. O curso clínico do linfoma folicular é bastante variável, havendo pouca concordância sobre qual a melhor estratégia de tratamento [109]. Em um estudo retrospectivo, DAVE *et al.* [89] analisaram 191 amostras de biópsias de pacientes com linfoma folicular não-tratados, usando arranjos de oligonucleotídeos. Quatro subgrupos de pacientes foram definidos com grande variância de sobrevida entre eles. Os genes identificados como mais informativos na definição dos subgrupos eram expressos por células imunes invasivas, e não pelas próprias células tumorais, evidenciando a importância da resposta imune na sobrevida dos pacientes desse tipo de câncer. Esta análise de expressão gênica permitiu a identificação de pacientes com bom prognóstico, que sobreviveram mais de 10 anos sem tratamento, e representam 75% dos pacientes com linfoma folicular.

Outra forma bastante heterogênea de câncer, no que diz respeito ao prognóstico clínico, é o cancer de mama. Aproximadamente 70% dos pacientes com doença negativa para os nodos axilares são curados por cirurgia, enquanto os 30% restantes apresentam doença agressiva. Os fatores prognósticos existentes são incapazes de diferenciar com precisão entre esses dois subtipos. VAN'T VEER *et al.* [96] usaram microarranjos de

DNA para analisar a expressão de aproximadamente 25000 genes de 78 pacientes com doença linfonodo-negativa abaixo de 55 anos de idade. Esses pesquisadores encontraram uma assinatura de expressão composta de 70 genes, que foi capaz de predizer o desenvolvimento de metástase no período de 5 anos após a cirurgia em 65 desses pacientes (acurácia predita de 83%). A aplicação do classificador prognóstico com 70 genes em um subconjunto independente de 19 pacientes de câncer de mama resultou em apenas duas classificações incorretas. Nenhum dos marcadores biológicos prognósticos bem estabelecidos para o câncer de mama como *ER*, *HER-2*, *uPA* ou *PAI-1* foram listados no conjunto de 70 genes preditores [96]. Dentre os genes selecionados como mais úteis na predição do câncer de mama, incluíram-se aqueles envolvidos no ciclo celular, transdução de sinal, invasão, metástase e angiogênese. Os fatores prognósticos incluiram *MMP-9*, *ciclina E2*, *RAB6B*, *MCM* e o receptor de *VEGF* (*FLT1*).

O perfil prognóstico desses 70 genes foi mais tarde testado em 295 pacientes atendidos em seqüencia no Instituto do Câncer Holandês [96]. Todos estes pacientes tinham doença em estádio I ou II e tinham menos de 53 anos de idade. Desses, 151 apresentavam doença linfonodo-negativo, enquanto que 144 apresentavam doença linfonodo-positiva. Dez dos 151 pacientes acometidos pela doença linfonodo-negativa e $122/144$ com doença linfonodo-positiva receberam terapia sistêmica adjuvante. Neste estudo, a probabilidade de permanecer livre de metástase distante durante 10 anos após a cirurgia foi de $85,2\%$, nos pacientes de melhor perfil prognóstico, e de $50,6\%$, naqueles de pior prognóstico. A razão de chances estimada para metástase distante entre os pacientes de pior prognóstico, e aqueles com melhor prognóstico, foi de $5,1$ (IC de 95%, $2,9$–$9,0$; $p < 0,001$). O perfil dos genes selecionados foi igualmente preditivo para paciente com a doença linfonodo-negativa, quanto para a forma linfonodo-positiva.

Em outro trabalho relacionado, também desenvolvido na Holanda, WANG *et al.* [91] identificaram 76 genes preditores do câncer de mama linfonodo-negativo. Neste estudo, que incluiu 286 pacientes (151 indivíduos na amostra de treinamento e 171 na amostra de validação), o perfil gênico foi um preditor mais importante que a idade, tamanho tumoral, estadiamento e estado do receptor de hormônio. Após 5 anos de acompanhamento, as diferenças em tempo livre de metástases e tempo de sobrevida entre os pacientes de melhor e pior prognóstico foram, respectivamente, de 40% e 27%. O classificador foi capaz, inclusive, de estimar o prognóstico de pacientes com tumor de 1–$2\,\mathrm{cm}$ de diâmetro, um subgrupo particularmente de difícil predição prognóstica. Diferente do

estudo de VAN'T VEER *et al.* [96], este estudo não incluiu pacientes que receberam terapia sistêmica adjuvante. Foi possível então diferenciar o impacto do prognóstico de forma isolada do possível fator de confundimento relacionado à terapia sistêmica.

Outros grupos mostraram ainda, que perfis de expressão gênicos podem diferenciar entre cânceres de mama agressivos ou indolentes [110, 92, 93, 94, 95]. Assim como em DLBCL, pouca ou nenhuma superposição existe entre os genes identificados pelos diferentes grupos [111]. Por exemplo, apenas 17 genes na lista dos 231 identificados por VAN'T VEER *et al.* [96] e 465 identificados por SORLIE *et al.* [92] eram comuns. Ainda, apenas dois genes se sobrepunham nos estudos de RAMASWAMY *et al.* [94] e SORLIE *et al.* [92], enquanto apenas três genes foram comuns entre os estudos de VAN'T VEER *et al.* [96] e WANG *et al.* [91].

O câncer de próstata é outra malignidade de prognóstico clínico particularmente heterogêneo. Com a introdução do rastreamento pelo PSA, o câncer de próstata está agora sendo detectado em estádios mais precoces. Enquanto a maioria dos cânceres de próstata é indolente, e nunca gera manifestações clínicas da doença, o problema principal na administração desses pacientes está na seleção daqueles que requerem tratamento agressivo. DHANASEKARAN *et al.* [97] compararam o perfil de expressão gênico entre tecidos prostáticos malignos e benignos. Um dos genes apresentando grande diferença de expressão entre esses dois tipos de tumor foi o da protease de serina, hepsina. A medida de hepsina em 334 cânceres de próstata de 78 homens com câncer de próstata localizado, mostrou que níveis baixos, ou ausentes, correlacionavam-se com a elevação do nível de PSA, sugerindo a presença de micrometástases.

SINGH *et al.* [98] também empregaram microarranjos de DNA na identificação prognóstica do câncer de próstata. Neste estudo, 21 pacientes foram selecionados para acompanhamento pós-cirúrgico. Oito desses pacientes foram identificados como tendo a forma recidivante da doença (definido como duas elevações consecutivas do PSA sérico), enquanto 13 ficaram livres da doença. Apesar da expressão de nenhum gene se correlacionar com o prognóstico, a assinatura de 5 genes foi preditiva, são eles: cromogranina, fator beta de crescimento de plaquetas, *HOXC6*, receptor de inositol trifosfato e sialotransferase.

A maior limitação dos dois estudos acima foi que o mesmo conjunto de dados clínicos foi usado tanto para treinamento quanto para validação. Contrastando com esses estudos, GLINSKY *et al.* [99] usaram um grupo de 21 pacientes com câncer de próstata

para descobrir os genes preditores do prognóstico, e um segundo grupo de 79 pacientes para validar o perfil preditor. A análise de sobrevida mostrou que a assinatura de 5 genes a partir de amostras de tumor obtidas após prostatectomia cirúrgica radical, era capaz de predizer o prognóstico de pacientes com estádio inicial da doença, tanto em pacientes com nível de PSA pré-operatório alto quanto baixo. Ao final, 88% dos pacientes com recorrência do câncer de próstata, em 1 ano após a terapia, foram classificados corretamente no grupo de prognóstico ruim.

MOCH *et al.* [100] estudaram, em 5184 clones de cDNA, os fatores prognósticos para o carcinoma de células renais empregando linhagens celulares *CRL-1933* e tecido renal normal. Entre os 89 genes diferencialmente expressos apontados, encontrava-se um que codificava para a vimentina, um filamento citoplasmático intermediário, cuja significância para prognóstico foi comprovada por teste imunocitoquímico.

ESCHRICH *et al.* [102] estudaram o estadiamento molecular para predição de sobrevida em pacientes com câncer de cólon e reto. Foi realizada a correlação entre a classificação de microarranjos de cDNA contendo 32 mil clones e a sobrevida, em 78 amostras de câncer de cólon e reto. O estadiamento molecular, baseado em uma assinatura de 43 genes (incluindo a osteopontina e neuregulina), obteve acurácia de 90% (sensibilidade de 93% e especificidade de 84%) na predição da sobrevivência em 36 meses, mostrando-se significativamente superior ($p < 0,001$) ao estadiamento clínico padrão atual (Sistema de Estadiamento de Dukes), particularmente para os pacientes com estádios B e C.

Outros estudos no qual os arranjos de DNA foram usados na predição do prognóstico incluem: carcinoma hepato celular [103], carcinoma escamoso de cabeça e pescoço [104], Linfoma (DLBCL) [88], câncer de mama [96], câncer de cólon e reto [101], câncer de pulmão [105], câncer esofágico [106] e blastoma medular [107].

Além das possibilidades da utilização de microarranjos de DNA citadas, ou seja, como auxílio diagnóstico e determinação de prognóstico, é crescente também sua utilização na obtenção de marcadores preditivos da resposta terapêutica (tab. 2.3). Os marcadores preditivos são freqüentemente necessários na oncologia, já que apenas a minoria dos pacientes com uma forma particular de câncer respondem à terapia convencional para ela adotada.

Na tentativa de desenvolver melhores marcadores preditivos para a terapia hormonal, MA *et al.* [20] investigaram o perfil de expressão de 60 pacientes com câncer de

Tabela 2.3: Exemplos de trabalhos usando microarranjos de DNA na determinação de marcadores preditivos da resposta terapêutica.

Malignidade	**Terapia**	**Publicação**
Câncer de mama	Quimioterapia (Tamoxifeno)	MA *et al.* [20]
Câncer de próstata	Terapia hormonal anti-andrógena	CHEN *et al.* [21]

mama receptor de estrogênio-positivos, ou *ER*-positivos, tratados com o adjuvante tamoxifeno. Usando um microarranjo de oligonucleotídeos de 22000 genes, 19 genes foram selecionados como diferencialmente expressos entre os pacientes hormônio-sensíveis e hormônio-resistentes. Após a microdissecção a laser das células tumorais, 9 genes foram considerados como diferencialmente expressos entre os pacientes sensíveis e resistentes à terapia hormonal. Apenas 2 genes foram identificados tanto no preparo a partir do tecido "total", como na microdissecção a laser, são eles, *HOXB13* e o receptor da interleucina *17B* (*IL17BR*). Análise com curvas ROC mostraram que os transcritos de *HOXB13* e *IL17BR* foram superiores na predição do prognóstico após tratamento com o adjuvante tamoxifeno que os marcadores em uso, ou seja, o receptor de estrogênio e o *PR* (receptor de progesterona).

Diferente da situação com o câncer de mama, não há atualmente a disponibilidade de nenhum marcador preditivo para a resposta à terapia hormonal no câncer de próstata. Na tentativa de identificar mecanismos de resistência à terapia hormonal anti-andrógena no câncer de próstata, CHEN *et al.* [21] investigaram o perfil de expressão gênico global de 7 pares xenográficos isogênicos hormônio-sensitivos e hormônio-resistentes. Dos 12599 genes testados (conjunto de sondas), apenas um gene, aquele para o receptor de androgênio (AR), estava diferencialmente expresso nos sete pares de tumor. Consistente com os achados dos microarranjos de DNA, a análise de *Western Blotting* mostrou que os cânceres refratários ao tratamento expressaram uma quantidade maior de proteína AR que seus pares. O próximo passo, segundo os autores, consiste na observação do comportamento do aumento da expressão de AR na resposta (sensibilidade/resistência) à terapia anti-andrógena.

Nos últimos anos, um número de pesquisadores usaram o perfil de expressão gênico para predizer a quimiossensibilidade (tab. 2.4), isto é, a sensibilidade à quimioterápicos usados em oncologia, usando linhagens celulares.

Tabela 2.4: Exemplos de trabalhos usando microarranjos de DNA para predizer a sensibilidade à quimioterápicos usados em oncologia via células de linhagem celulares tumorais.

Nº de Linhagens celulares	**Nº de drogas anti-câncer**	**Publicação**
60	118	SCHERF *et al.* [112]
39	55	DAN *et al.* [113]

Em um dos primeiros estudos, SCHERF *et al.* [112] correlacionaram a expressão de 1376 genes com a resposta à 118 drogas anti-câncer, em 60 linhagens celulares humanas. Essas células de linhagem tiveram origem em diversas doenças malignas, incluindo câncer de cólon, rim, ovário, mama, próstata e pulmão, e foram extensivamente caracterizadas em função da resposta farmacológica a múltiplos agentes anti-câncer. A aglomeração das linhagens celulares em função do seu perfil de expressão apresentou resultados distintos daqueles obtidos pela aglomeração das linhagens celulares em função da sensibilidade à drogas. Em particular, os autores identificaram exemplos de interações gene-droga, onde a alteração no nível de expressão gênico era consistente com o mecanismo de ação da droga. Por exemplo, os autores encontraram uma significante correlação entre dihidropirimidina desidrogenase (*DPD*) que converte a droga *5-FU* em dihidrofluorouracil (*DHFU*), e a resposta à *5-FU*.

Em estudo similar, DAN *et al.* [113] usaram microarranjos de DNA na investigação da sensibilidade de 39 linhagens celulares de cânceres humanos a 55 agentes anti-câncer. Esses autores encontraram, que enquanto alguns genes correlacionavam-se conjuntamente com várias classes de drogas, outros genes correlacionavam-se apenas com drogas específicas, de mecanismo de ação similar. Por exemplo, a aldose redutase, que cataliza a conversão de glicose a sorbitol, estava associada com a sensibilidade a 24 drogas. Por outro lado, o gene anti-apoptótico, survivina, se correlacionou apenas com a resposta a pirimidinas.

Até hoje, poucos trabalhos usaram perfis de expressão gênicos na predição da *sensibilidade clínica* a drogas anti-câncer (tab. 2.5).

Tabela 2.5: Exemplos de trabalhos usando microarranjos de DNA na predição da *sensibilidade clínica* a drogas anti-câncer.

Malignidade	**Quimioterapia**	**Publicação**
Câncer de mama	Docetaxel	CHANG *et al.* [114]
	paclitaxel, *5-FU*, doxorubicina e ciclofosfamida (*T/FAC*)	AYERS *et al.* [115]
	tamoxifen	JANSEN *et al.* [116]
Leucemia linfoblástica aguda (ALL)	Daunorubicina, vincristine, prednisolone ou asparaginase	HOLLEMAN *et al.* [117]
	Imatinib (*Glivec*)	HOFMANN *et al.* [118]

Em um relato preliminar, em 24 pacientes com câncer de mama, CHANG *et al.* [114] encontraram 92 genes diferencialmente expressos em tumores oriundos de pacientes sensíveis e resistentes à terapia com a droga docetaxel. Em uma análise de validação-cruzada do tipo *leave-one-out* (LOOCV), $10/11$ tumores sensíveis e $11/13$ tumores resistentes foram corretamente classificados (acurácia de 88%). A assinatura de 92 genes apresentou valores preditivos positivos e negativos, respectivamente, de 92% e 83%. Os resultados foram validados em um conjunto independente de apenas 6 pacientes. Os tumores sensíveis apresentaram elevada expressão dos genes envolvidos no ciclo celular, citoesqueleto, adesão, transporte de proteínas e apoptose, enquanto os tumores resistentes apresentaram aumento de expressão de genes de transcrição e sinal de tradução.

AYERS *et al.* [115] também usaram microarranjos de DNA na identificação de genes associados com a resposta à terapia química em pacientes com câncer de mama. Neste estudo, a quimioterapia usada foi a seqüencia paclitaxel e *5-FU* $+$ doxorubicina $+$ ciclofosfamida (*T/FAC*), o número de pacientes investigado foi de 42 (24 usados na descoberta e 18 na validação independente) e o rótulo de classe foi dado pela resposta patológica completa. Usando-se uma assinatura de 74 genes, foi encontrada uma acurácia de predição de 78% $(14/18)$. A razão de resposta esperada ao *T/FAC* para o tratamento de pacientes não selecionados com câncer de mama é de 28%.

Uma assinatura de 44 genes para predição de resistência ao tamoxifeno, em pacientes com câncer de mama, foi definida por JANSEN *et al.* [116], baseada em 112 tumores *ER*-positivos analisados via microarranjos de cDNA de aproximadamente 18000 genes. A assinatura não melhorou a predição da resposta, quando comparada as medidas clínicas em uso, mas melhorou a predição da progressão do tumor durante o tratamento.

Outra malignidade na qual a tecnologia dos microarranjos de DNA foi usada na identificação de marcadores de terapias é a ALL. Aproximadamente, 80% das crianças com ALL são curadas com quimioterapia. As razões para o fracasso do tratamento nas restantes 20% ainda são desconhecidas. Na tentativa de abordar esse problema de resistência, HOLLEMAN *et al.* [117] investigaram células de 173 crianças com ALL para a sensibilidade, *in vitro*, a daunorubicina, vincristine, prednisolone ou asparaginase. O perfil de expressão gênico foi então usado na seleção de genes diferencialmente expressos em células de ALL droga-resistentes e droga-sensíveis. No total, 172 genes foram encontrados como diferencialmente expressos nas linhagens de células B de leucemias sensíveis e resistentes. Destes, 22 genes estavam incluídos no caso da daunorubicina, 59 para vincristine, 42 para prednisolone e 54 para asparaginase. A aglomeração hierárquica acertou o estado de resistência em $86/105$ casos para daunorubicina, $84/104$ para vincristine, $66/75$ para prednisolona e $83/106$ com relação a asparaginase. Um perfil de expressão gênico para resistência aos quatro agentes estava associado com um aumento significativo de recidiva da doença. A medida de resistência combinada também foi preditiva do prognóstico do tratamento usando análise multivariada, que incluía as variáveis: idade do paciente, subtipo genético da ALL, linhagem da ALL e contagem de leucócitos ao diagnóstico. Estes resultados foram confirmados em uma população independente de pacientes tratados de forma igual àqueles 173 pacientes originais.

Em outro estudo em ALL, HOFMANN *et al.* [118] usaram microarranjos de DNA na identificação de genes conferindo resistência ao inibidor da tirosina quinase, imatinib (*Glivec*) em pacientes com ALL. Este estudo contou com 19 pacientes adultos com ALL Philadelphia-positivos incluídos na fase II dos testes de segurança e eficiência do imatinib. Usando 95 genes, os autores foram capazes de separar todos os casos imatinib-sensíveis dos casos resistentes. Dentre os genes mais expressos nas células resistentes de ALL estavam a tirosina quinase de Bruton e duas ATP sintetases (*ATP5A1* e *ATP5C1*). Genes com decréscimo de expressão nessas células incluíram o gene pro-apoptótico *BAK1* e o gene de controle de ciclo celular *15INK4B*.

Capítulo 3

Seleção de Atributos em Microarranjos de DNA por Algoritmos Genéticos

3.1 Algoritmos Genéticos

Pode-se entender a evolução como um processo de otimização, que não tem a perfeição como propósito final, mas que é capaz de descobrir soluções altamente precisas e funcionais para um problema imposto por um ambiente a um organismo [119]. Os *algoritmos genéticos* (GAs) são métodos de busca estocásticos, que mimetizam a evolução biológica natural. Os GAs operam em uma população de soluções potenciais, aplicando-se o princípio da sobrevivência do mais apto, afim de produzir aproximações cada vez melhores de uma solução.

Os primeiros conceitos usados, no que hoje denomina-se GA, foram apresentados por HOLAND [120] no começo dos anos 60. Na visão deste, a característica básica de um sistema adaptativo, natural e robusto, era o uso bem sucedido da competição e inovação, provendo a habilidade de responder dinamicamente a eventos não antecipados e à mudanças ambientais. Modelos simples de evolução biológica pareciam facilmente capturar essas idéias, via noções de sobrevivência do mais apto e produção continuada de descendentes.

Uma descrição informal de um GA genérico poderia ser feita como segue (fig. 3.1). A população é inicializada arbitrariamente, e evolui no sentido de regiões cada vez melhores do espaço de busca através de processos estocásticos de *seleção* (que são determi-

nísticos em alguns algoritmos), *mutação*, e *recombinação* (que é completamente omitida em outros algoritmos). O ambiente (objetivo traçado pelo processo de otimização/busca) estipula um valor qualitativo (*índice de aptidão*) dos pontos de busca, e o processo de seleção favorece aqueles indivíduos mais aptos.

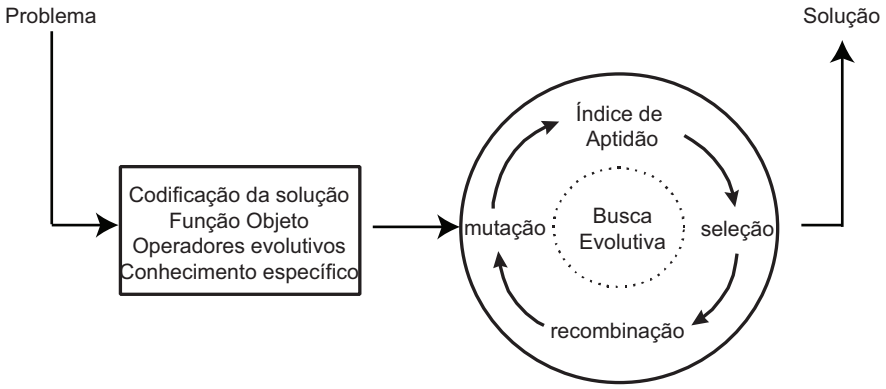

Figura 3.1: Diagrama da solução de problemas usando-se algoritmos genéticos. Havendo um problema, que se pretende otimizar via GAs, há primeiro a necessidade de codificação do problema na forma de uma/várias populações (algoritmo de uma população ou multi-populacional), de tamanho (número de indivíduos em cada população) definido, de cromossomos binários (`bits` 0s e 1s), dos operadores de seleção (índice de adaptação), recombinação (ou *crossover*) e re-inserção, que conjuntamente determinarão o algoritmo a ser empregado, assim como da função de avaliação (ou objetivo)—associada ao critério de otimização (geralmente uma função de erro, com/sem restrições e penalizações)—e o critério de parada do algoritmo (número de ciclos ou valor da função de avaliação abaixo de um valor pré-determinado). Ao final, há a necessidade de se decodificar os cromossomos da última geração, afim de se obter a solução do problema.

Os *GAs canônicos*, como descritos e analisados por HOLLAND [41] e GOLDBERG [42], tomam alguns termos emprestados da genética populacional. Assim, as estruturas individuais são denominadas de *cromossomos*. Elas são genótipos, que são manipulados pelos GAs. A rotina de avaliação decodifica essas estruturas em alguma estrutura fenotípica e a atribui um *índice de aptidão*. Essa representação fenotípica determina o algoritmo a ser utilizado, ou seja, os operadores evolutivos admissíveis.

Na representação por valores binários, utilizada nos GAs canônicos, os cromossomos são normalmente seqüencias de `bits` de comprimento fixo. O valor em cada *locus* (`bit`) nesta seqüência é denominado *alelo*. Algumas vezes os *loci* individuais

(sub-seqüencias de `bits`) são também chamados de genes. Em outras ocasiões, genes são combinações de alelos, que apresentam algum significado fenotípico, como, por exemplo, parâmetros dos operadores evolutivos.

A noção de *avaliação* e *aptidão* é algumas vezes inter-cambiável. É útil distinguir entre *função objetivo*/*objeto*, ou *função de avaliação*, e a *função de aptidão* usadas em GAs. A função objetivo, fornece uma medida de desempenho em relação a um conjunto particular de parâmetros, enquanto a função de aptidão transforma essa medida de desempenho em uma alocação de oportunidade de reprodução. A avaliação de uma seqüencia, representando um conjunto de parâmetros, é independente da avaliação de qualquer outra seqüencia. A aptidão dessa seqüencia, entretanto, é sempre definida em relação a outros indivíduos da mesma população.

Pode-se ver a execução de um algoritmo genético como um processo de dois estágios. Este começa com a *população atual*. A seleção é aplicada à essa população afim de criar uma *população intermediária*. Então, a recombinação e a mutação são aplicadas na população intermediária para criar uma nova população. Este ciclo é repetido até que o critério de parada seja alcançado. O processo, partindo de uma população atual para a próxima população, constitui uma *geração* na execução do GA (fig. 3.2). Os GAs operam em populações (soluções múltiplas), ao invés de em indivíduos (soluções únicas). O processo de busca/otimização em GAs ocorre de forma paralela, tal qual o processo de pensamento humano.

O *operador de seleção* em GAs enfatiza a regra de probabilidade de sobrevivência, misturada com uma chance, dependente da aptidão, de se obter diferentes casais que produzirão um número maior ou menor de filhos/descendentes. O primeiro passo na seleção é a atribuição do índice de aptidão a cada indivíduo da população. Cada indivíduo no conjunto de seleção recebe uma probabilidade de reprodução dependendo do valor de sua função objetivo, assim como do valor da função objetivo de outros indivíduos do conjunto de seleção. Este valor é usado na seleção, propriamente dita, no passo seguinte.

Há duas formas de se atribuir a aptidão [121, 122]: a *atribuição da aptidão baseada na ordenação* e a *atribuição proporcional à aptidão*. Na primeira [123, 124], os indivíduos da população atual são ordenados de acordo com sua avaliação, e a probabilidade de seleção de cada indivíduo é uma função dessa ordem, ou seja, considerando-se N_{ind} o número de indivíduos na população atual, Pos a posição de um certo indivíduo nessa população (o indivíduo de menor valor da função objetivo tem $Pos = 1$, e o de

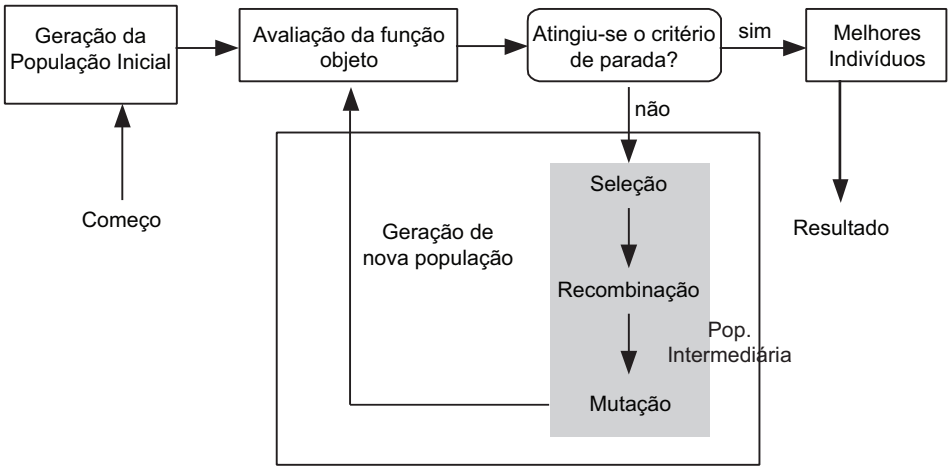

Figura 3.2: Estrutura de um Algoritmo genético de uma única população. Primeiro, gera-se uma população inicial (1ª população atual), que, determinada a função objetivo, será por essa avaliada. Se não for atendido o critério de parada, seleciona-se os indivíduos que comporão a população intermediária (região sombreada), que sofrerá recombinação (*crossover*), mutação e re-inserção na nova população atual (final de 1 geração, ou ciclo do GA). Essa última será novamente avaliada e, caso não atenda ao critério de parada, será inserida em um novo ciclo de otimização até, que se atinja o objetivo de otimização (parada). Ao final, os melhores indivíduos são decodificados, gerando um conjunto de soluções ótimas para o problema de otimização.

maior valor $Pos = N_{ind}$) e SP a pressão seletiva, a aptidão (Fit) atribuída a cada indivíduo é calculada como:

$$Fit(Pos) = 2 - SP + 2(SP - 1)\frac{Pos - 1}{N_{ind} - 1}, \tag{3.1}$$

sendo admissíveis, para o caso linear, valores de SP entre 1 e 2. No caso não-linear se introduz uma distribuição não-linear. Este pode ser calculado como segue:

$$Fit(Pos) = \frac{N_{ind} X^{Pos-1}}{\sum_{i=1}^{N_{ind}} X^{i-1}}, \tag{3.2}$$

onde X é computado pela raiz do polinômio:

$$(SP - 1)X^{N_{ind}-1} + SPX^{N_{ind}-2} + \ldots + SPX + SP = 0, \tag{3.3}$$

admitindo valores de SP entre 1 e $(N_{ind} - 2)$.

Segundo BÄCK e HOFFMEISTER [124] e WHITLEY [123], a atribuição de aptidão baseada na ordenação é mais robusta que a proporcional à aptidão. Estes afirmam, que a primeira suplanta os problemas de escalonamento da atribuição proporcional, ou seja, estagnação nos casos onde a pressão seletiva é muito pequena, ou convergência prematura para mínimos locais, onde a seleção ocasiona uma rápida restrição do espaço de busca. Na aptidão baseada na ordenação, a taxa de reprodução é limitada, de forma que nenhum indivíduo gera um número excessivo de descendentes. A ordenação introduz um escalonamento uniforme na população e permite o controle simples e eficiente da pressão seletiva.

Atribuído o valor de aptidão a cada indivíduo da população atual, a seleção propriamente dita é computada. Os indivíduos da população atual são selecionados de acordo com sua aptidão por meio de um dos seguintes algoritmos: *seleção pela roda-da-roleta*, *amostrador estocástico universal*, *seleção truncada* e *seleção por torneio*.

O esquema mais simples de seleção é a seleção pela roda-da-roleta, também chamado de *amostrador estocástico com reposição* [121] (fig. 3.3). Este é um algoritmo estocástico, onde os indivíduos são mapeados em segmentos contíguos de uma reta, de forma que o segmento de cada indivíduo seja proporcional, em tamanho, a sua aptidão. Um número aleatório (distribuição uniforme entre 0 e 1) é gerado, e o indivíduo que tiver o número gerado em seu segmento é selecionado. O processo é repetido até que o número desejado de indivíduos seja obtido.

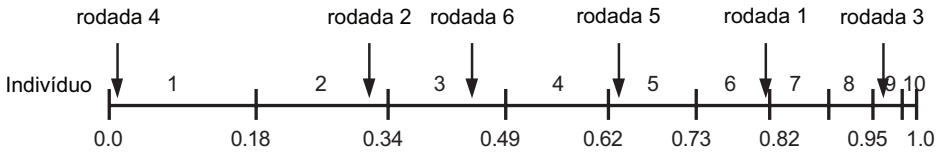

Figura 3.3: Seleção pela roda-da-roleta.

No amostrador estocástico universal [121], assim como acontece na seleção pela roda-da-roleta, os indivíduos são mapeados em segmentos contíguos de uma reta proporcionais à aptidão de cada um. Como mostrado na figura 3.4, ponteiros eqüidistantes são posicionados sobre o segmento linear em quantidade igual ao número de indivíduos a ser selecionado. Considerando-se $N_{Pointer}$ o número de indivíduos a ser selecionados, a distância entre os ponteiros é de $1/N_{Pointer}$, e a posição do primeiro ponteiro é dada por um número aleatório gerado no intervalo $[0, 1/N_{Pointer}]$.

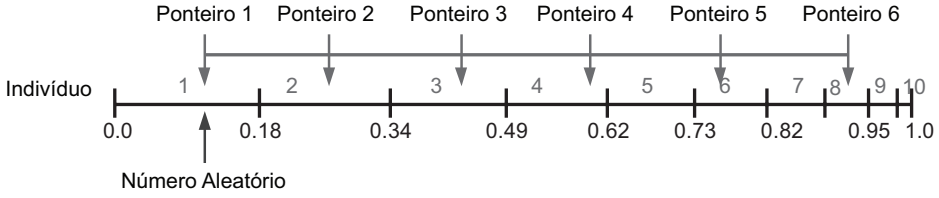

Figura 3.4: Amostrador estocástico universal.

Na seleção truncada [122], apenas os melhores indivíduos, em função de suas aptidões, são selecionados da população atual. Esses indivíduos selecionados produzem descendentes de forma aleatória e uniforme. O parâmetro do operador de seleção truncada é o ponto-de-corte para a seleção, $Trunc$, que indica a proporção da população a ser selecionada como população intermediária, e assume valores entre 50% e 10%. Indivíduos abaixo do ponto-de-corte são excluídos da população intermediária.

Na seleção por torneio [125] um subconjunto de indivíduos é selecionado aleatoriamente na população atual, e o melhor indivíduo desse subconjunto é selecionado para a população intermediária. Este processo é repetido com freqüência igual ao número de indivíduos a ser escolhido para a população intermediária. Estes indivíduos selecionados produzem descendentes de forma aleatória e uniforme. O parâmetro para a seleção por torneio, $Tour$, determina o número de indivíduos selecionado (tamanho do subconjunto) em cada passo, admitindo valores entre 2 e N_{Tour}, onde N_{Tour} é o número de indivíduos a formarem a população intermediária.

Todos os GAs (sejam eles de população única ou múltipla) funcionam combinando o operador de seleção a um mecanismo que combine informação dos pares de indivíduos da população intermediária (recombinação), e a outro que produza variação (mutação) nesses indivíduos recombinados.

Por motivos históricos, a recombinação em GAs canônicos, que utilizam a representação binária, é conhecida como *crossover*. A idéia do *crossover* é a de que, dados dois indivíduos que são altamente aptos, mas o são por razões distintas, a recombinação deles dará origem a pelo menos um novo indivíduo que combine as melhores características de cada um. Como desconhecemos, *a priori*, essas características, o melhor que se pode fazer é recombiná-las ao acaso. O operador de *crossover* trata estas características como blocos de construção espalhados pela população, e tenta recombiná-los em indivíduos melhores via recombinação. Algumas vezes o *crossover* irá combinar as piores características de dois indivíduos em seu descendente, que neste caso, não continuará a ser selecionado.

Durante o *crossover* apenas partes dos indivíduos são trocadas. O número de pontos de cruzamento/troca diferencia os operadores. No *crossover de ponto único* (fig. 3.5), uma única posição de troca $k \in \{1, 2, \ldots, N_{var} - 1\}$, onde N_{var} é o número de `bits` de um indivíduo, é selecionada de forma aleatória e uniforme, dada uma razão de recombinação, que varia normalmente no intervalo $[0,6, 1, 0]$. A seguir, os `bits` a partir desse ponto são trocados entre os indivíduos, produzindo dois indivíduos. Já no *crossover* de ponto duplo, como o próprio nome diz, duas posições de troca são determinadas de forma aleatória e uniforme, e os `bits` entre esses pontos são trocados entre os indivíduos, gerando dois novos indivíduos.

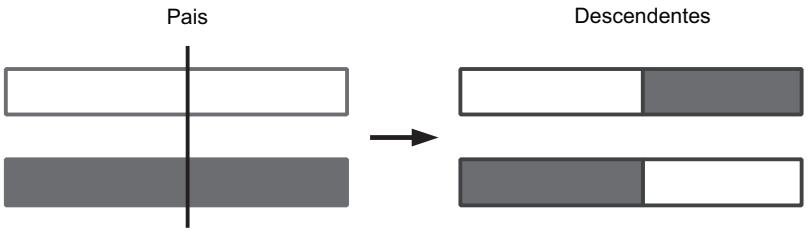

Figura 3.5: *Crossover* de ponto único.

O *crossover* de ponto único e o *crossover* de ponto duplo podem ser considerados como casos especiais do método geral denominado de *crossover* de pontos múltiplos (fig. 3.6). Neste, m posições de troca $k_i \in \{1, 2, \ldots, N_{var} - 1\}$, para $i = \{1, \ldots, m\}$, são

escolhidas aleatoriamente de forma uniforme e sem repetição, e, então, ordenadas (ordem crescente). A seguir, os `bits` entre pontos sucessivos de *crossover* são trocados entre os dois indivíduos selecionados, gerando dois novos indivíduos. Os `bits` entre o primeiro `bit` e o primeiro ponto de *crossover* não são trocados entre os indivíduos.

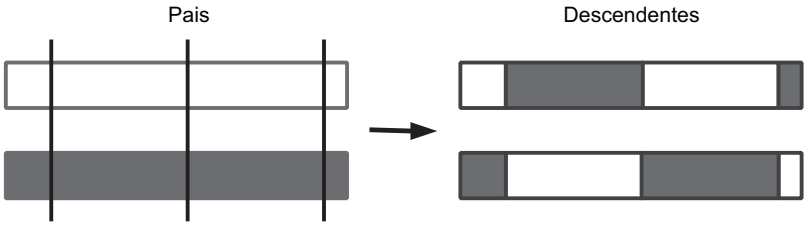

Figura 3.6: *Crossover* de pontos múltiplos.

A idéia do operador de *crossover* de pontos múltiplos, é que as partes do cromossomo que mais contribuem para o desempenho de um indivíduo particular, não necessariamente estão contidas em seqüências de `bits` adjacentes [126]. Além disso, a natureza destrutiva do operador *crossover* de pontos múltiplos parece fortalecer a exploração no espaço de busca, evitando a convergência precoce (pontos mínimos locais), pela seleção sucessiva de indivíduos altamente aptos, tornando assim, mais robusta a otimização [127].

O operador *crossover* uniforme leva o número de pontos de *crossover* ao extremo, usando uma decisão aleatória de fazer ou não a troca de informação *bit-a-bit* entre os indivíduos selecionados [128]. Para cada posição, o indivíduo que contribui com sua variável (0 ou 1) ao indivíduo recombinado é escolhido de forma uniforme e aleatória, com a seguinte probabilidade:

$$Var_{i_{recomb}} = Var_{i_{Ind_1}} a_i + Var_{i_{Ind_2}} (1 - a_i), \quad \text{para } i \in \{1, 2, \ldots, N_{var}\}, \tag{3.4}$$

onde $a_i \in \{0, 1\}$ é escolhido de forma uniforme e aleatória. Por exemplo, na recombinação das seqüencias de `bits`, ou cromossomos, 0000 e 1111, podemos ver, que estas seqüencias estão conectadas por um conjunto de caminhos mínimos em um hipercubo de dimensão quatro. Se alterarmos um único `bit` na representação binária de origem, estaremos um passo mais próximo do destino. Pode-se ver na figura 3.7, que alterando-se um `bit` indica um movimento para cima ou para baixo no gráfico.

O operador *crossover* uniforme, assim como o *Crossover* de pontos múltiplos, diminui o *bias* associado ao comprimento da representação binária usada na codificação de um dado parâmetro. SPEARS e DE JONG [129] demonstraram a forma como o operador

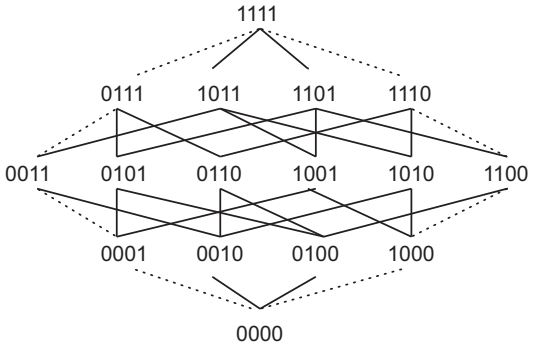

Figura 3.7: O operador *crossover* uniforme. Ilustração de caminhos em um espaço de dimensão igual a 4. Os operadores de *crossover* uni-dimensional e multi-dimensional podem gerar descendentes ao longo dos caminhos pontilhados nas bordas do gráfico.

crossover uniforme pode ser parametrizado pela aplicação de uma razão de probabilidade da troca de `bits`. Esse parâmetro extra pode ser usado no controle da amplitude de alteração durante a recombinação, sem com isso, introduzir *bias* em relação ao tamanho da representação usada.

Além dos operadores citados para a representação binária das variáveis, há ainda os operadores *shuffle crossover* [130] e *crossover with reduced surrogate* [126]. O primeiro seleciona um posição de troca (como no *crossover* de ponto único), mas antes que a troca de `bits` ocorra, estes são embaralhados em ambos os indivíduos selecionados. No segundo, o operador de *crossover* restringe o universo de indivíduos recombinantes gerados, impondo que esses indivíduos sejam obrigatoriamente novos indivíduos, ou seja, diferentes dos outros indivíduos recombinantes gerados anteriormente na mesma geração. Não há, até então, uma teoria clara, nem evidência empírica, para que se decida qual operador de *crossover* utilizar.

Uma vez selecionados os indivíduos da população intermediária, e recombinados, pelos operadores apresentados, o GA impõe aos indivíduos dessa população intermediária um mecanismo de variação, afim de tornar mais diverso o universo de busca e, conseqüentemente, menor a probabilidade de convergência precoce para pontos mínimos locais, O mecanismo de variação mais conhecido é a *mutação*.

Para representações binárias de indivíduos, a mutação significa a inversão do valor de um `bit`, já que cada *locus* apresenta apenas duas possibilidades (0 ou 1). Para cada

indivíduo, são aplicadas alterações aleatórias com baixa probabilidade [1] (*razão de mutação* ou *probabilidade de mutação*). A posição da variável a ser alterada na seqüencia cromossômica é escolhida, normalmente, de forma uniforme e aleatória (*passos de mutação*). Duas possibilidades são possíveis quando da definição do operador de mutação. Na primeira, os dois parâmetros (passos de mutação e razão de mutação) são constantes durante o processo evolutivo, enquanto, na segunda, os parâmetros se adaptam entre mutações consecutivas.

Apesar de muitos GAs usarem conjuntamente a mutação e a recombinação (*crossover*), para muitos problemas de otimização, a utilização de GAs com mutação na ausência de *crossover* pode ser bastante eficaz [131].

Uma vez que os indivíduos da população intermediária foram produzidos por seleção, recombinação e mutação dos indivíduos da população atual, cabe agora decidir quais indivíduos comporão a população atual da nova geração. Isso se dá por operadores conhecidos como *operadores de reinserção*, que atuam inserindo ou removendo indivíduos das populações atual e intermediária, afim de compor a população dessa nova geração. Por exemplo, se há um número menor de indivíduos na população intermediária, quando comparado ao tamanho da população atual, então indivíduos da população atual devem ser *reinseridos* na população intermediária. Similarmente, se nem todos os indivíduos da população intermediária devem ser aproveitados, ou ainda, se há a geração de um número de indivíduos nessa população maior do que aquele da população atual, então um esquema de *remoção* deve ser usado na determinação de quais indivíduos devem ser perpetuados na nova geração.

O método de seleção aplicado anteriormente à população atual, é novamente empregado no esquema de reinserção. Há diferentes esquemas de reinserção global. Pode-se, por exemplo, gerar um número de indivíduos na população intermediária igual ao número de indivíduos na população atual, e substituí-los todos (*reinserção simples*); pode-se gerar um número menor de indivíduos na população intermediária, e nela inserir indivíduos da população atual, selecionados de forma uniforme e aleatória (*reinserção aleatória*); pode-se gerar um número menor de indivíduos na população intermediária, e nela simplesmente inserir indivíduos da população atual de maior aptidão (*reinserção elitista*); ou ainda, pode-se gerar um número de indivíduos na população intermediária maior do que

[1] Geralmente, usam-se valores próximos a 1×10^{-3} *per bit*.

o da população atual, e então inserir apenas os indivíduos da população intermediária de maior aptidão (*reinserção baseada na aptidão*).

Na reinserção simples, cada indivíduo vive apenas uma geração. É, então, muito provável, que indivíduos muito bons sejam substituídos sem que tenham gerado descendentes ainda melhores, o que resulta em perda de informação. A reinserção elitista, quando combinada com a reinserção baseada na aptidão, previne essa perda de informação (fig. 3.8). A cada geração, um dado número de indivíduos menos aptos é substituído pelo mesmo número de indivíduos da população intermediária de melhor aptidão. O esquema de reinserção baseada na aptidão implementa a seleção truncada entre os indivíduos da população intermediária antes de reinserí-los na nova população. Assim, os indivíduos mais aptos na população atual podem sobreviver por muitas gerações sem que se perca a inserção de indivíduos das populações intermediárias, como acontece na aplicação isolada da reinserção baseada na aptidão. Não se controla se os indivíduos da população atual são melhores ou piores que seus substitutos da população intermediária. Como indivíduos da população intermediária menos aptos podem substituir indivíduos mais aptos da população atual, a aptidão média da população pode decrescer entre gerações consecutivas, entretanto, se os indivíduos da população intermediária inseridos são muito pouco aptos, espera-se que eles sejam substituídos por novos indivíduos de populações intermediárias futuras, não havendo assim, ao passar das gerações, uma perda de convergência de otimização.

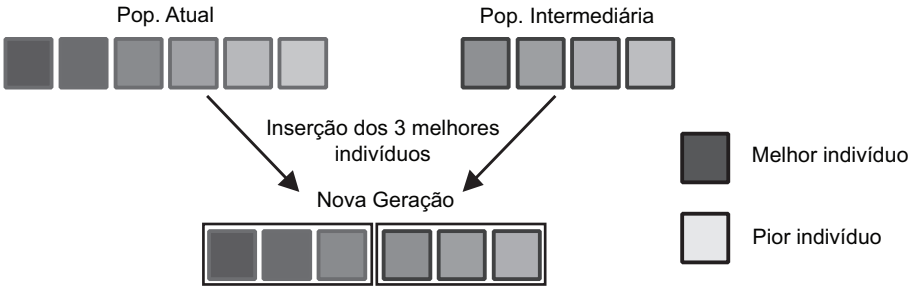

Figura 3.8: Esquema de reinserção combinando os operadores de reinserção elitista e reinserção baseada na aptidão.

Uma vez gerada a nova população atual (nova geração), cada indivíduo é avaliado. Calcula-se o valor da função objetivo (função de avaliação), que é o propósito final da otimização. Se o critério de parada/otimização for atendido, por exemplo, a variação do valor da função de erro por um número pré-determinado de gerações for igual a zero, para-

se o algoritmo, caso contrário, aplicam-se novamente os operadores de seleção, mutação, recombinação e reinserção, até que esse critério seja atendido (fig. 3.2).

3.2 Algoritmos Genéticos de Objetivos Múltiplos

Problemas reais geralmente apresentam critérios múltiplos (múltiplos objetivos). Baseado na comparação desses vários objetivos, muitas vezes conflitantes, pode ser tomada a decisão sobre a superioridade de um indivíduo sobre outros e, conseqüentemente, sua preservação no processo evolutivo. Formalmente, a otimização de objetivos múltiplos lida com a minimização simultânea de $NObj$ objetivos, com funções f_r, onde $r = 1, \ldots, NObj$. Os valores de f_r são determinados por funções objetivos (fenótipo), que, por sua vez, são dependentes das variáveis dos indivíduos (genótipo). Um problema clássico pode ser usado como motivação para as considerações a seguir. Quando brinquedos são produzidos, os custos da produção devem ser mantidos baixos e o tempo de produção deve ser o menor possível. Quando há a diminuição do custo de produção, por exemplo, pela terceirização de etapas da produção, há um aumento do tempo final de produção. Em contrapartida, se todas as etapas de produção são realizadas localmente, preferencialmente em uma única planta, diminui-se o tempo de produção às custas de um aumento de custo do produto final. Os objetivos da otimização, custo (f_1) e tempo (f_2) de produção (fig. 3.9), servem de avaliação para cada solução.

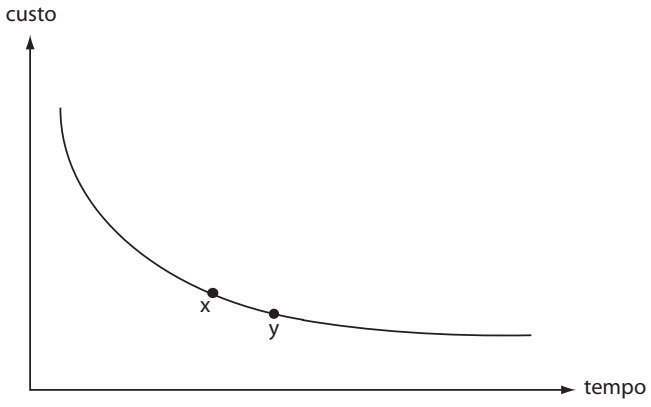

Figura 3.9: A relação entre custo e tempo.

Os Algoritmos Genéticos de Objetivos Múltiplos (MOGA) aparecem como uma solução natural à procura por uma população de pontos com propriedades desejadas, uma vez que os algoritmos genéticos são processos paralelos que buscam não uma, mais soluções diversas no espaço de busca. A otimização de objetivos múltiplos por algoritmos genéticos pode ser dividida em três grupos [132, 133]: (1) Abordagens agregadoras; (2) Abordagens não-Pareto baseadas em populações; e (3) Abordagens Pareto-próximas.

Nas abordagens agregadoras, os objetivos são numericamente combinados em um único objetivo a ser otimizado. O método mais simples e usado dessa abordagem é o de soma-ponderada (*weighted-sum approach*). Suas principais vantagens são a fácil implementação e a possibilidade de ser utilizado em qualquer algoritmo otimizador. Suas desvantagens, entretanto, são que a escolha dos pesos atribuídos aos objetivos são problema-dependente e, que as soluções localizadas em regiões côncavas do espaço de busca—regiões onde ocorrem conflitos entre os objetivos do problema—não podem ser encontradas (fig. 3.10). Outros algoritmos incluem o *minimax* e o vetor de alvo (*target vector approach*) [132, 133].

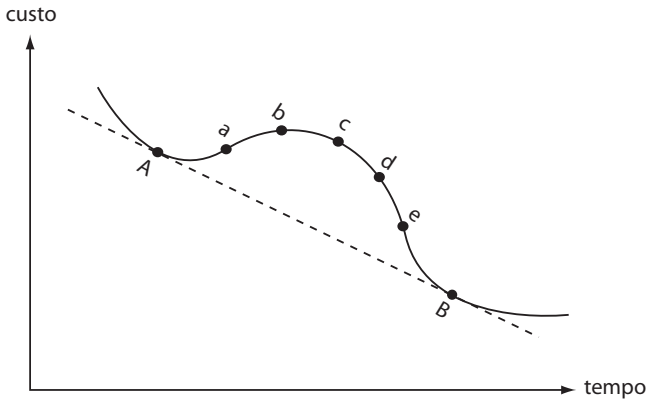

Figura 3.10: Algumas soluções Pareto-ótimas (**a**, **b**, **c**, **d** e **e**) estão "escondidas" em um espaço objetivo não-convexo.

Nas abordagens não-Pareto baseadas em populações, objetivos diferentes afetam a seleção ou exclusão de diferentes partes da população de forma ordenada. Incluem-se nesse grupo a abordagem lexicográfica (*lexicographic approach*), a abordagem VEGA, e a abordagem da ordenação pela mediana (*median-rank approach*) [132, 133].

As abordagens Pareto-próximas buscam soluções mais próximas o possível do conjunto Pareto-ótimo (Pareto *front*), encontrando uma coleção de soluções possíveis. A abordagem mais importante desse grupo é a de ordenamento Pareto (*Pareto ranking*). Diferente das abordagens anteriores, não há parâmetros (pesos) problema-dependentes, mas há a necessidade de uma estratégia de ordenamento (seleção) baseada no conceito de dominância (Pareto-ótimas) e, ainda, de estratégias que garantam a manutenção da diversidade do espaço de busca. A seguir, apresentaremos alguns conceitos importantes à compreensão dessas abordagens.

A superioridade de uma solução sobre outras pode ser decidida pela comparação de duas ou mais soluções. Esta comparação pode ser feita seguindo as definições de inferioridade, superioridade, ou a não-inferioridade. Um vetor de soluções $\mathbf{f1} = [f1_1, f1_2, \ldots, f1_{NObj}]$ é dito inferior ao vetor $\mathbf{f2} = [f2_1, f2_2, \ldots, f2_{NObj}]$, onde $r = 1, \ldots, NObj$, se $\mathbf{f2}$ é parcialmente menor que $\mathbf{f1}$, $\mathbf{f2}\, p < \mathbf{f1}$ (fig. 3.11), ou seja,

$$\forall r = 1, \ldots, NObj, \quad \mathbf{f2_r} \leq \mathbf{f1_r} \qquad \wedge \qquad \exists r = 1, \ldots, NObj \, : \, \mathbf{f2_r} < \mathbf{f1_r}. \tag{3.5}$$

Da mesma forma, o vetor $\mathbf{f1}$ é dito superior a $\mathbf{f2}$, se $\mathbf{f2}$ é inferior a $\mathbf{f1}$. Os mesmos são ditos não-inferiores, se $\mathbf{f2}$ não é nem inferior nem superior a $\mathbf{f1}$.

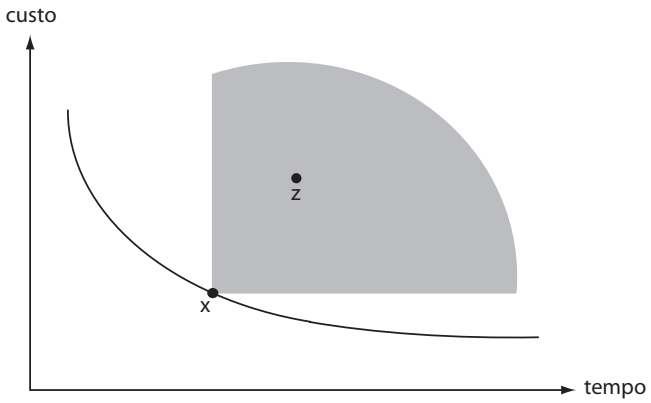

Figura 3.11: Soluções dominantes. A solução $\mathbf{x}$ gera uma área de dominação (sombreada). Qualquer solução nessa área, como $\mathbf{z}$, é uma solução dominada.

Se a solução $\mathbf{f2}$ é parcialmente menor que $\mathbf{f1}$, há a indicação de que $\mathbf{f1}$ domina $\mathbf{f2}$. No exemplo anterior, isso significa que se o custo ou o tempo para a solução $\mathbf{f1}$ forem menores do que para a solução $\mathbf{f2}$, então a solução $\mathbf{f1}$ é superior à solução $\mathbf{f2}$. Seria

suficiente se um dos dois valores fossem iguais para as duas soluções (ex.: custos iguais) e o outro valor fosse menor para **f1** (ex.: menor tempo).

Do conceito de soluções não-dominadas vem o de conjunto de soluções Pareto-ótimas. Esse pode ser definido como o conjunto encontrado de soluções não-dominadas do espaço de busca. Entretanto, um conjunto de soluções não-dominadas pode conter tanto soluções Pareto-ótimas, quanto soluções não-Pareto-ótimas. A tarefa de qualquer algoritmo para encontrar conjuntos Pareto-ótimos é retornar, a cada geração, um conjunto de soluções não-dominadas mais aptas.

Dentre as estratégias de seleção por ordenamento Pareto (fig. 3.12), três se destacam. Na primeira, proposta por GOLDBERG [42], todos os indivíduos não-dominados na população recebem valor de aptidão 1, e são removidos. Um novo conjunto de indivíduos não-dominados é identificado e esses recebem valor de aptidão 2. O processo continua até que toda a população seja ordenada. Na segunda, proposta por SRINIVAS e DEB [134] os indivíduos não-dominados na população recebem valor de aptidão igual ao número de indivíduos na população e são removidos da população, o mesmo acontecendo com os conjuntos não-dominados restantes. Na terceira, proposta por FONSECA [135], os indivíduos recebem valor de aptidão correspondente ao número de indivíduos na população que os dominam. Nas três estratégias os valores de aptidão são corrigidos afim de se garantir a diversidade das soluções, como visto a seguir. Após atribuídos os valores de aptidão, a seleção é feita usando operadores como *Amostrador Estocástico Universal*, *Seleção pela roda-da-roleta*, ou *Seleção por torneio*.

Cada solução no conjunto Pareto-ótimo obtido pode ser interessante. Os MOGAs são mais que a procura por soluções Pareto-ótimas. Para produzirem soluções aceitáveis, os métodos de solução de objetivos múltiplos necessitam incorporar preferências humanas (informações de alto-nível). Usando-se tanto a seleção baseada na ordenação de FONSECA [135], SRINIVAS e DEB [134], como a de GOLDBERG [42], as soluções Pareto-ótimas são equivalentes, ou seja, recebem o mesmo valor inicial de aptidão. Apesar de todas as soluções Pareto-ótimas serem superiores àquelas por elas dominadas, nem todas as soluções constituem soluções aceitáveis do problema. Usando-se o exemplo anterior da fábrica de brinquedos, pode-se ter uma solução onde o custo de produção seja igual a zero e o tempo de produção seja infinito. Nenhuma outra solução pode produzir brinquedos com custo menor e, conseqüentemente, essa seleção não pode ser dominada, pertencendo ao conjunto Pareto-ótimo de soluções. Em outro exemplo extremo, soluções

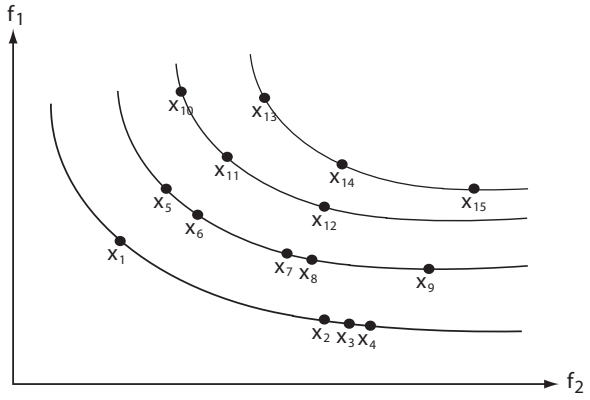

Figura 3.12: Estratégias de seleção baseadas no conceito de dominância. Várias soluções são classificadas em diferentes frontes não-dominadas.

podem produzir brinquedos em um tempo muito curto, mesmo às expensas de custos extravagantes. É evidente que nenhuma das opções é desejada, apesar de pertencerem ao conjunto de soluções não-dominadas.

FONSECA [135] introduziu a possibilidade de se incluir valores desejados (*goals*), definidos pelo usuário, para as funções objetivos da otimização. Assim, uma solução é aceitável apenas quando os valores pré-definidos são alcançados. Esse procedimento é conhecido como *método das inequações* ou *goal programming*, onde os valores desejados são inseridos como inequações. No exemplo da produção de brinquedos, inequações indicando um limite superior para custo e tempo de produção podem ser determinados. O resultado não é aceito até que ambos os valores sejam simultaneamente alcançados, ou seja, até que sejam suficientemente próximos. Quando o método das inequações é adotado, a comparação entre duas soluções aumenta em complexidade, sendo assumidos os seguintes pressupostos a partir das definições dadas na eq. 3.5: (1) Se a solução $\mathbf{f1}$ não atender aos valores desejados e a solução $\mathbf{f1}$ for $p <$ que a solução $\mathbf{f2}$, então a solução $\mathbf{f2}$ é preferida; (2) Se a solução $\mathbf{f1}$ atender a todos os valores desejados e a solução $\mathbf{f1}$ for $p <$ que a solução $\mathbf{f2}$, ou a solução $\mathbf{f2}$ não atender aos valores desejados, então a solução $\mathbf{f2}$ é preferida; e, (3) Se a solução $\mathbf{f1}$ atender a alguns dos valores desejados e o vetor dado pelo subconjunto de soluções de $\mathbf{f1}$, que atendem aos valores desejados, for $p <$, ou igual, ao vetor dos mesmos objetivos para a solução $\mathbf{f2}$ e o vetor dos demais objetivos da

solução $\mathbf{f1}$ for $p <$ que o vetor dos mesmos objetivos da solução $\mathbf{f2}$, ou ainda, esse vetor de $\mathbf{f2}$ atender aos valores desejados, então a solução $\mathbf{f2}$ é preferida.

Um algoritmo genético normalmente converge para uma única solução. Esse processo é conhecido como *genetic drift*, ou perda da diversidade genética. Métodos evolutivos devem manter a diversidade genética ao longo do processo evolutivo, prevenindo assim a convergência prematura do algoritmo. A manutenção da diversidade do espaço de busca é obtida pelo favorecimento de soluções de maior diversidade no espaço de decisão (genótipo), ou ainda, no espaço dos objetivos (fenótipo). Nem sempre, e principalmente em problemas não-lineares complexos, a diversidade no espaço dos objetivos implica em diversidade no espaço de decisão. Com isso pretende-se uma melhor aproximação entre o conjunto de soluções não-dominadas e, o objetivo do algoritmo, o conjunto Pareto-ótimo (fig. 3.13).

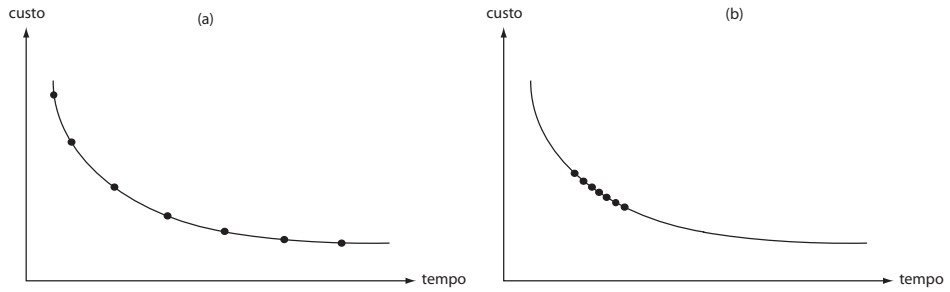

Figura 3.13: Manutenção da diversidade do espaço de busca. Dois conjuntos de frontes Pareto-ótimas. A diversidade em (a) é "maior" que a diversidade em (b).

Os métodos conhecidos de preservação da diversidade de soluções em algoritmos genéticos podem ser divididos em [136]: métodos de nicho (*Niching methods*) e métodos de não-nicho. Entre os métodos de não-nicho, o *restricted mating* é o mais comum. Neste, dois indivíduos da população intermediária podem se recombinar apenas se estiverem a uma certa distância, restrita por uma constante, um do outro. Já os métodos de nicho podem ser divididos em duas categorias: *Fitness sharing* e *crowding*.

O *Fitness sharing*, como introduzido por GOLDBERG e RICHARDSON [137], é um mecanismo de escalonamento da aptidão, que altera apenas a etapa de seleção do GA. A idéia do *Fitness sharing* é a de que o número de indivíduos que pode residir em qualquer porção do espaço de busca (objetivo ou decisão) é limitado pela aptidão da porção do espaço que ocupa. Disso resulta a alocação ótima de indivíduos no espaço. O número de

indivíduos residindo em picos (maximização), ou vales (minimização), será proporcional ao quanto é alto, ou baixo, aquele pico (vale).

O *Fitness sharing* trabalha pela diminuição da aptidão de cada indivíduo na população, proporcional ao número de indivíduos similares na população. Especificamente, o valor de aptidão compartilhada de cada indivíduo, F', é igual à sua aptidão inicial, F, dividida pela contagem no seu nicho. A contagem do nicho individual é a soma dos valores da função de compartilhamento, $sh(.)$, entre si e cada indivíduo na população (incluindo a si próprio). A aptidão compartilhada $F'(i)$ de uma solução i é dada pela seguinte equação:

$$F'(i) = \frac{F(i)}{\sum_{j=1}^{\mu} sh(d(i,j))}. \tag{3.6}$$

A função de compartilhamento é uma função da distância entre dois elementos da população; ela retorna 1 se os elementos são idênticos, 0 se são mais distantes que um ponto-de-corte de similaridade, e um valor nesse intervalo para níveis intermediários de dissimilaridade. O ponto-de-corte para a similaridade é controlada por uma constante σ_{share}. Uma função de compartilhamento típica é:

$$sh(d) = \begin{cases} 1 - (d/\sigma_{share})^{\alpha}, & \text{se } d < \sigma_{share}; \\ 0, & \text{caso contrário.} \end{cases} \tag{3.7}$$

onde α é uma constante que regula a forma da função de compartilhamento.

O *Niching* GA pode aplicar métricas de distância tanto genotípicas quando fenotípicas. No *sharing* genotípico, a função de distância $d(.)$ é simplesmente a distância de Hamming entre duas seqüencias, ou seja, o número de `bits` diferentes entre duas seqüencias. No *sharing* fenotípico, a função de distância é definida de acordo com o problema, sendo a distância Euclidiana comumente empregada para valores contínuos.

As técnicas de *crowding* inserem novos elementos na população pela substituição de elementos similares. Assim como nas técnicas de *sharing* para determinar similaridade, os métodos de *crowding* utilizam métricas de distância tanto no genótipo, quanto no fenótipo. Os métodos de *crowding* tendem a espalhar os indivíduos entre os picos mais proeminentes (maximização) do espaço de busca.

3.3 Seleção de Atributos em Dados de Microarranjo de DNA

Independente da tecnologia usada em sua fabricação, dada a grande dimensão (número de genes por lâmina) dos microarranjos, uma etapa de grande importância é a seleção de atributos (variáveis), ou seja, o procura dos genes, ou seus padrões/perfis de expressão gênicos, relevantes para o experimento realizado. Isto, por si só, é uma procura desafiadora, considerando que encontrar um classificador usando n atributos (perfis gênicos) de um total de N atributos possibilita $D = \mathrm{N}!\,/(\mathrm{n}!\,(\mathrm{N}-\mathrm{n})!)$ possibilidades diferentes de seleção. Por exemplo, para $n = 10$ e $N = 3000$ há $D = 1,6 \times 10^{28}$ seleções possíveis.

Quando o questionamento experimental envolve a análise do *padrão de expressão gênica*, ou seja, quando se está interessado nos genes que apresentam expressão diferenciada (indução ou repressão) em um arranjo (experimentos de comparação de duas amostras—ex. controle $\times$ amostra de interesse), há diversas opções disponíveis, importando sobretudo, se há replicação do experimento. Na ausência de replicação, utiliza-se o *método da razão*. Neste, utilizam-se os dados de expressão e_{ij}, selecionando-se genes que apresentam esse valor absoluto acima de um ponto-de-corte pré-determinado (geralmente $1,5$ ou 2), o que indicaria um aumento/diminuição de expressão da amostra de interesse em relação ao controle (ou, a outra amostra de interesse) na ordem de $2,25$ a 4 maior/menor. Em experimentos com replicação, há a possibilidade de utilização de métodos estatísticos mais confiáveis [138, 139, 140]. Exemplos desses métodos incluem: teste t-*student* [138, 139], ANOVA [141, 142], testagens não-paramétricas [82, 140], e Bayesianas [143, 28, 144, 145, 146, 147].

Já quando o questionamento experimental envolve a análise das amostras, ou do *padrão de expressão amostral*, pretende-se saber quais genes melhor representam, ou discriminam, padrões amostrais s_j conhecidos e rotulados (ex. dois fenótipos distintos de leucemia). Dado o escopo do trabalho, nos concentraremos nestes métodos.

Métodos de seleção de variáveis (atributos ou características) supervisionados, ou seja, quando se conhece os rótulos das amostras pesquisadas, são generalizados em dois tipos [24]: *métodos de filtragem* (*filter methods*) e *métodos envelopados* (*wrapper methods*). A diferença essencial entre esses dois métodos é, que no método envelopado utiliza-se o algoritmo treinado (classificador/função discriminante) para a discriminação

de amostras, enquanto no método de filtragem a seleção de variáveis é independente do algoritmo treinado.

O método envelopado avalia o desempenho do classificador treinado em cada subconjunto do conjunto original dos atributos/variáveis testado, via uma função de erro, escolhendo, ao final do processo de seleção, o subconjunto de variáveis de melhor desempenho. Já o método de filtragem, não utiliza o classificador treinado e, ao invés de comparar o desempenho do classificador com diferentes subconjuntos de variáveis, tenta predizer os subconjuntos mais promissores utilizando estatísticas calculadas nas distribuições empíricas das classes do problema. Para a redução de genes em microarranjos, esses métodos são usualmente denominados, respectivamente, de *abordagens de ordenação de subconjuntos de genes* e *abordagens de ordenação individual de genes*.

A *abordagens de ordenação individual de genes* é a mais usada em análise de microarranjos de DNA. Nesta abordagem, para cada gene (atributo/característica) é medida a correlação com a classe (tipo tecidual, prognóstico, resposta a um quimioterápico), segundo algum critério medido (ex. distância Euclidiana, índice de correlação de Pearson, Informação Mútua, etc.). Após a seleção dos genes, que melhor atendem ao critério estipulado, é feita a ordenação dos mesmos. A característica principal da ordenação gênica é que esta baseia-se na correlação individual, de cada perfil gênico, com as distribuições de classe, ou seja, não se explora a correlação entre os perfis gênicos. Genes selecionados dessa forma, apresentam individualmente alta correlação com a classe, mas, quando associados a outros genes, podem não gerar o melhor desempenho do classificador. Desde que cada gene é considerado individualmente, alguns genes podem conter a mesma informação de correlação, introduzindo assim, redundância no processo seletivo. Genes que são complementares entre si na determinação dos rótulos de classe podem não ser selecionados, ocorrendo, com freqüência, quando estes não exibem, individualmente, alta correlação com a classe.

O método proposto por GOLUB *et al.* [17], conhecido como correlação GS [2], propõe uma métrica de correlação que mede a separação relativa de classes (problemas binários), produzida pelos valores de expressão de um gene. Esta métrica favorece genes que apresentam grandes variações de valores de expressão médios entre classes, e pequenas variações dos valores de expressão dentro da classe. Para um gene i, sejam $\bar{x}_{1i}$ e $\bar{x}_{2i}$ os valores de expressão médios de i nas duas classes, e sejam $\tilde{s}_{1i}$ e $\tilde{s}_{2i}$ os desvios-padrão

[2] Sigla adotada segundo as iniciais de seus autores Golub, T. R. e Slonim, D. K..

dos valores de expressão de i nas classes 1 e 2, a métrica de seleção gênica é dada por $P(i) = (\bar{x}_{1i} - \bar{x}_{2i})/(\tilde{s}_{1i} + \tilde{s}_{2i})$. O valor $P(.)$ de cada gene é medido, os genes agrupados em valores positivos e negativos, e ordenados segundo seus valores absolutos. Os $k/2$ genes de maior valor em cada grupo são selecionados, onde k é o número de genes a ser selecionado.

DUDOIT *et al.* [25], propuseram uma modificação na correlação GS. Segundo estes, a utilização de $\tilde{s}_{1i}+\tilde{s}_{2i}$ seria um modo incomum, e errado, de se calcular o erro-padrão da diferença dos valores de expressão médios do gene i entre as duas classes. Estes então definem a função $P(i)$ com base na razão entre a soma dos quadrados entre grupos ($BSS(i)$) e a soma dos quadrados dentro do grupo ($WSS(i)$), após normalização dos microarranjos para média 0 e desvio-padrão 1 (normalização, onde a distância entre as amostras de mRNA poderiam, segundo os autores, ser medidas pela distância Euclidiana). Para um gene i, essa razão é:

$$\frac{BSS(i)}{WSS(i)} = \frac{\sum_j \sum_k I(y_j = k)(\bar{x}_{ki} - \bar{x}_{.i})^2}{\sum_j \sum_k I(y_j = k)(x_{ji} - \bar{x}_{ki})^2}, \tag{3.8}$$

onde $\bar{x}_{.i}$ indica o nível de expressão médio do i-ésimo gene entre todas as m-amostras, e $\bar{x}_{ki}$ indica o nível de expressão médio do i-ésimo gene entre as amostras pertencentes à classe k, indicados pela função indicadora $I(.)$, $I(y_j = k)$.

GOH *et al.* [26] propuseram um método híbrido, que combina o coeficiente de correlação de Pearson e a correlação GS. Primeiro, são calculados os coeficientes de correlação de Pearson entre os n perfis de expressão gênicos do arranjo. Genes altamente correlacionados entre si são considerados dependentes, ou corregulados, e formam agrupamentos. A ordenação pela correlação GS é aplicada aos melhores indivíduos de cada agrupamento (de maior correlação), afim de se obter aqueles mais correlacionados com as classes conhecidas do problema, ou seja, mais discriminantes. Os k genes com maior valor de $P(.)$ no método GS são selecionados e, via um algoritmo incremental, começando com 1 gene, são aplicados no treinamento de uma função de classificação.

PARK *et al.* [27] introduziram uma abordagem de ordenação individual de genes, baseada em uma ordenação não-paramétrica, denominada desordem, robusta a valores extremos (*outliers*). O valor atribuído a cada gene é definido como o menor número de trocas de dígitos consecutivos necessária a uma perfeita separação, ou desordem. Sendo $e_i = [e_{i,1}, e_{i,2}, \ldots, e_{i,n}]$ o vetor indicativo do perfil de expressão gênico do gene i em

uma ordem crescente de 1 até n, e $c = [y_1, y_2, \ldots, y_n]$, $y_j \in \{0, 1\}$, os rótulos de classe correspondendo a e_i, o valor d atribuído ao perfil e_i é definido como:

$$d = \sum_{i \in I(y_l=1)} \sum_{k \in I(y_j=0)} \operatorname{sgn}(y_l - y_i), \tag{3.9}$$

onde, $I(.)$ é uma função indicadora de classe e $\operatorname{sgn}(.)$ é uma função sinal, retornando 0, se $d \leq 0$ e 1 se a condição não for atendida.

Outra abordagens de ordenação individual de genes, proposta por KELLER *et al.* [28], implementa uma medida de verossimilhança na seleção gênica para classificadores empíricos de Bayes (Naïve Bayes). Os genes são selecionados por uma função de ordenamento de máxima verossimilhança, LIK, computada para cada gene em cada classe. Considerando-se o caso binário (duas classes), onde C_1^i indica a distribuição da classe 1 para o i-ésimo gene, e C_2^i indica a distribuição da classe 2 para esse mesmo gene, as medidas de $\log$-verossimilhança podem ser definidas como: $LIK_{1\rightarrow 2} = \log p(C_1^i|X_1) - \log p(C_2^i|X_1)$ e $LIK_{2\rightarrow 1} = \log p(C_2^i|X_2) - \log p(C_1^i|X_2)$, onde X_1 e X_2 referem-se às tuplas [3] de treinamento nas classes 1 e 2, respectivamente. Um gene com poder discriminatório de classe, idealmente terá valores de LIK altos em ambas medidas, significando que os valores de expressão desse gene nas tuplas pertencentes à classe 1 votarão para classe 1, e àquelas pertencentes à classe 2, votarão para classe 2. Na prática, foi demonstrado que era difícil encontrar genes para os quais as duas medidas fossem muito maiores que 0. Assim, KELLER *et al.* [28] decidiram ordenar os genes para cada medida, individualmente, com a restrição de que a segunda medida deveria ser, obrigatoriamente, maior que 0. Após computar as medidas de LIK para cada gene, os genes em cada conjunto são ordenados de acordo com seu valor de LIK naquele grupo. Os $k/2$ genes com maior valor são selecionados de cada conjunto, onde k é o número de genes a ser selecionado.

BEN-DOR *et al.* [29] tratou o problema da ordenação individual de genes pela ordenação baseada em uma medida de relevância para cada gene. A intuição por trás dessa medida é a de que genes informativos apresentam valores bastante dissimilares entre as duas classes do problema, e que poder-se-ia separá-las por um ponto de corte. Os autores então definiram um ordenamento, intitulado TNoM, baseado no número de erros de classificação cometidos, em uma amostra de teste, por um classificador baseado em

[3] Unidades matemáticas composta pelos dados amostrais, $\mathbf{x}$, e seu respectivo rótulo, y. Representação: $(\mathbf{x}, y)$.

regras definidas a partir desse ponto de corte t, calculado para cada gene e_i, pertencentes à amostra de treinamento, definido como:

$$TNoM(e_i) = \min_{c,t} Err(c, t|e_i), \tag{3.10}$$

onde,

$$Err(c, t|e_i) = \sum_j 1\{l_j \neq \text{sgn}(c(e_i j - t_i))\}, \tag{3.11}$$

com parâmetros c indicando a classe predita, t indicando o ponto de corte de um perfil de expressão gênico e_i e l_j o rótulo de classe da amostra j, $j \in \{-1, +1\}$.

A segunda categoria de seleção de genes em microarranjos de DNA é a *abordagens de ordenação de grupos de genes*. Métodos dessa categoria procuram um subconjunto de genes, que gerem conjuntamente, a melhor função de discriminação de classes. A idéia é remover genes individualmente e monitorar o efeito da remoção desse gene no valor esperado de erro do classificador treinado.

Uma forma conhecida de se selecionar variáveis baseia-se no cálculo exaustivo da função de erro para classificadores construídos com todos os subconjuntos possíveis de variáveis. Esta estratégia é impraticável no caso dos microarranjos, dada a grande dimensão (número de genes), característica dos dados. Uma heurística popular de busca, mais eficiente, é denominada *eliminação retrógrada seqüencial*. Nesse método, começa-se com o conjunto total de genes, ou ainda, um subconjunto menor selecionado por uma abordagem de ordenação individual de genes e, a partir desse conjunto/subconjunto, remove-se, seqüencialmente, um gene por vez. A cada iteração, o gene removido é aquele que causa a menor redução no desempenho do classificador entre todos os candidatos à remoção.

GUYON *et al.* [30] propuseram a utilização da eliminação retrógrada seqüencial, chamada por estes de *eliminação recursiva de atributos* (RFE) na seleção de genes. Esta método consiste dos passos: (1) Treinar o classificador de Máquina de Vetor de Suporte (*Support Vector Machine*; SVM) (otimizar os pesos w_i); (2) Ordenamento de todas os perfis gênicos e_i pelo valor otimizado de seus pesos w_i; e (3) Remover o gene com menor valor de w_i. Do processo de eliminação (SVM–RFE), constatou-se, que houve a tentativa de reter o grupo de genes com maior poder de classificação, e que os genes selecionados ao final não eram, necessariamente, aqueles mais relevantes individualmente. GUYON *et al.* [30] encontraram, que a abordagem de seleção de genes com SVM–RFE gerava

melhores resultados para a classificação em dados de expressão de malignidades humanas, que a abordagem de seleção individual de genes GS, mas o tempo de computação era muito grande. GUYON *et al.* [30] propuseram, a seguir, a remoção de grupos de genes a cada iteração, ao invés de remover um a um até que o número de genes fosse tão pequeno como algumas centenas e, só então, começar a eliminar genes individualmente.

FURLANELLO *et al.* [148] propuseram uma modificação na seleção com RFE [30], denominada E–RFE (*Entropy-based Recursive Feature Elimination*), onde grupos de genes pouco interessantes são eliminados de acordo com a entropia da distribuição de pesos de um classificador SVM treinado. Sem perda de desempenho de classificação, o E–RFE mostrou-se mais eficiente (mais rápido), quando comparado tanto com a versão original do RFE, como quando comparado com a versão mais rápida, que admite a remoção de grupos de genes a cada iteração. Já CHO *et al.* [149] propuseram a utilização da análise discriminante de Fisher com artifício da função de núcleo (*Kernel Fisher Discriminant Analysis*; KFDA) como função de classificação ao invés da SVM, na seleção retrógrada de perfis gênicos.

3.4 Seleção de Atributos em Microarranjos de DNA por Algoritmos Genéticos

Nos últimos anos, diversas aplicações de algoritmos genéticos (GAs) em classificação de microarranjos de DNA foram relatadas, onde, com poucas exceções, os GAs utilizados apresentaram pouca variação do GA canônico [41, 42]. A maioria desses estudos empregou o GA para a seleção ótima de genes e, para isso, usaram outro método para a classificação em si.

LI *et al.* [48] desenvolveram um método, denominado GA/kNN, onde o GA foi usado para a seleção de um subconjunto de genes de uma matriz de expressão gênica. Os genes foram classificados usando o algoritmo de vizinhos próximos (kNN), com $k = 3$. Ao invés de utilizar os genes selecionados em uma única simulação (ronda) do GA, o método GA/kNN seleciona os genes mais freqüentes de um grande conjunto de simulações (20000–40000), e treina o classificador kNN final baseado nesses genes, usando a distância Euclidiana como métrica. A medida de ajuste do GA (função objetivo) empregada, foi simplesmente a razão de amostras (perfis de expressão gênicos) corretamente classificadas

no conjunto de treino. Como os GAs são algoritmos estocásticos, simulações diferentes geram normalmente diferentes conjuntos de genes selecionados. A coerência da abordagem reside no fato de genes mais capazes de gerar classificações corretas irão aparecer no subconjunto selecionado com maior freqüência, sendo mais adequados à inclusão no classificador final.

O método foi aplicado ao conjunto de leucemia aguda [17], com o objetivo de diferenciar entre duas classes de leucemia aguda: linfóide (ALL) e mielóide (AML). Após a filtragem para a remoção de genes com baixos níveis de expressão, 5545 genes permaneceram, com medidas em 72 amostras, das quais 47 pertencentes à ALL e 25 à AML. O conjunto de dados foi então dividido em 38 amostras de treinamento e 34 amostras de teste (validação). O resultado demonstrou que o classificador kNN final, baseado nos 50 genes mais freqüentemente selecionados em 40000 tentativas, pôde classificar corretamente 33 das 34 amostras de teste usando a regra da maioria (duas de três) para o classificador kNN. É de interesse, que nenhum dos 40000 classificadores individuais obteve o mesmo desempenho. O melhor desempenho de classificadores individuais foi de 32 acertos em 34 amostras. LIU *et al.* [49] usaram um GA multi-populacional (paralelo) para a seleção gênica, combinado com o método de classificação proposto por GOLUB *et al.* [17], intitulado *weighted voting method*. Para o mesmo conjunto de dados, um classificador com 29 genes obteve uma acurácia de 88% (30 das 34 amostras).

LI *et al.* [48] relatam outra aplicação do método GA/kNN, onde o método foi aplicado aos dados de câncer de cólon reportados por ALON *et al.* [57]. Neste conjunto de dados, as 62 amostras (40 amostras tumorais e 22 normais) foram divididas em um conjunto de treinamento com 42 amostras e um conjunto de teste com 20 amostras. O melhor resultado relatado foi encontrado para classificadores usando entre 25–110 dos genes mais freqüentemente selecionados em tentativas individuais. LI *et al.* [48] realizou um estudo sistemático do desempenho do GA/kNN. Em tentativas envolvendo cromossomos curtos, com 5–10 genes, foi encontrado que uma lista pequena de genes era muito freqüente em muitas tentativas. O aumento do tamanho do cromossomo resultou em um padrão de seleção gênico mais estável, apesar de também aumentar o tempo de geração.

WAHDE e SZALLASI [150] também estudaram classificadores usando o menor número possível de atributos. Contrastando com métodos anteriores, o GA proposto não apenas selecionou genes a serem usados no classificador, mas também gerou o classificador, ou seja, otimizou os parâmetros do classificador. Os classificadores gerados eram

lineares, com ponto de corte, e são aplicáveis à classificação binária. Assim como em DEUTSCH [151], também foi usada uma medida de similaridade próxima à de ordenação $TNoM$ [29]. Durante o GA, os índices dos genes eram modificados como resultado de mutações. Ao invés de utilizarem um corte baseado na ordenação, como em [151], WAHDE e SZALLASI [150] permitiram, no princípio, a seleção de qualquer gene para inclusão no classificador, com preferência dada aos genes melhor ordenados. O método foi aplicado aos dados de câncer de mama publicado por VAN'T VEER *et al.* [96], que contém 5277 perfis de expressão gênicos em 97 amostras, das quais, 78 compuseram a amostra de treinamento, enquanto 19 foram utilizadas como teste. Usando apenas 7 genes, os autores encontraram classificadores com acurácia de treinamento de $97,4\%$ e de teste superior a $89,5\%$, sendo superior a publicações anteriores no mesmo conjunto de dados. Com apenas 4 genes, os autores obtiveram acurácia de treinamento de $94,9\%$ e superior a $84,2\%$ na amostra de teste.

Essas propostas de emprego de GAs na seleção gênica foram ampliadas por OOI e TAN [50], LIU *et al.* [45], LIN *et al.* [51]JIRAPECH-UMPAI e AITKEN [46] e DEUTSCH [151], que empregaram esses algoritmos na seleção gênica em problemas de classificação multi-classes.

OOI e TAN [50] apresentaram a seleção de perfis gênicos em combinação com o classificador Bayesiano de máxima verossimilhança (MLHD), denominado GA/MLHD. Neste, a medida de seleção foi escolhida como $f = A - (E_C + E_I)$, onde A é uma constante suficientemente grande para manter o valor de seleção não-negativo, E_C é o erro de validação cruzada na amostra de treinamento, e E_I é o erro obtido no conjunto de teste. O método GA/MLHD foi aplicado no conjunto de dados NCI60, contendo 64 linhagens de células cancerosas (das quais foram usadas 61), que foram divididas em nove classes baseadas no sítio de origem do câncer. De um total de várias centenas de genes, 1000 foram selecionados baseado no desvio padrão dos perfis de expressão gênicos. O método obteve $85,37\%$ de acurácia de classificação (medida pelo erro E_C), com um classificador de 13 genes. Entretanto, LIU *et al.* [45] afirmam que o erro de classificação foi forjado por OOI e TAN [50] e, que a verdadeira acurácia de classificação do GA/MLHD foi de apenas $70,73\%$.

LIN *et al.* [51] combinaram um GA com a classificação baseada na estatística de silhueta, usando a mesma medida de seleção de OOI e TAN [50]. Na aplicação do método proposto no conjunto de dados de NCI60, LIN *et al.* relataram uma acurácia de $90,3\%$

(i.e., um pouco maior que a encontrada por LIU e IBA [49]). LIN *et al.* também revisaram o desempenho de várias medidas de distância para a estatística de silhueta, e encontraram que a medida de Pearson era superior às demais.

JIRAPECH-UMPAI e AITKEN [46] propuseram a utilização de um GA combinado com o classificador kNN com tamanho populacional de $\{10, 30, 50\}$ indivíduos, e subconjuntos de genes de tamanho $\{30, 50\}$. Após 100 gerações. O melhor resultado médio, entre as 10 simulações, para os dados de leucemia [17], foi de $98,42\%$ (acurácia de classificação dada pela LOOCV), com tamanho populacional de 50 e número de genes de 50, mas com erro médio de teste de apenas $72,64\%$. Foram testados 6 diferentes métodos para a filtragem inicial do número de genes para um subconjunto de tamanho igual a 100. Usando o método de ordenação pela correlação GS [17], a média obtida em 10 simulações com GA de tamanho populacional também de 10 foi de $98,24\%$. A metodologia foi aplicada aos dados de classificação multi-classes NCI60 [152], e o melhor classificador, com GA de tamanho populacional de 30 e com 30 genes selecionados, obteve acurácia de teste de $76,23\%$, usando o mesmo método de filtragem.

DEUTSCH [151] considerou o problema de minimizar o tamanho dos classificadores, com o propósito de encontrar a menor assinatura gênica, resultando em uma redução de custo para aplicação clínica. No método desenvolvido pelos autores, denominado *genetic evolution of subsets of expressed sequences* (GESSES), um filtro (similar à ordenação TNoM, introduzida por BEN-DOR *et al.* [29]) é usado na formação do subconjunto inicial da matriz de expressão gênica. O algoritmo kNN, com $k = 1$, é então usado na classificação. A medida de adaptação baseia-se na acurácia LOOCV do classificador, mas adiciona um segundo componente com propósito de maximizar a separação de classes. O método GESSES é inicializado com um consórcio de classificadores, e remove, ou adiciona, genes até que o desempenho ótimo seja alcançado. Aplicando o método aos dados multi-classes de KHAN *et al.* [18], foram encontrados classificadores com número pequeno de genes (até 10), que conseguiram classificação perfeita das 20 amostras de teste. O número ideal de genes ($\pm$ desvio-padrão) relatado para esse conjunto de dados foi de 12 ± 2.

FRÖHLICH *et al.* [43], PENG *et al.* [44], LIU *et al.* [45], SOUZA e CARVALHO [40] e HUERTA *et al.* [47] propuseram a seleção de genes em dados de expressão, com duas ou mais classes, combinando algoritmos genéticos e Máquinas de Vetores de Suporte.

No trabalho de FRÖHLICH *et al.* [43] foram apresentados GAs especiais, intitulados GAR2W2, GAJH e GAAcc, onde a função objetivo era, respectivamente, vinculada ao limite do erro de generalização de VAPINIK [34], ao limite do erro de generalização de Jaakkola-Haussler [153], e ao erro de generalização estimado pela validação cruzada de k dobras. Para agregar os objetivos de menor erro de generalização e número de genes do classificador treinado, os autores propuseram a função objetivo de soma ponderada dada por $1 + 0,001 \times \frac{m}{d}$, onde m indicava o número de genes utilizados pelo classificador e d indicava o número total de genes nos conjuntos de dados de câncer de cólon [57] e no conjunto multi-classes (5 classes) de esporulação do fungo *S. cerevisiae* (Laboratórios Brown, http://cmgm.stanford.edu/pbrown/yeastchip.html). Nesses GAs, foram também evoluídos no cromossomo genes responsáveis pela codificação do parâmetro de regularização C, ou de margem *soft* ($0,001$, $0,01$, $0,1$, 1, 10, 100, 1000) e número de genes (20, 50, 100, 250, 500, 1000). Quando o número de genes e o parâmetro de regularização eram fixados antes do processo evolutivo, o GAAcc apresentou erro de generalização significativamente pior que o GAR2W2 e o GAJH, mas semelhantes ao erro obtido com algoritmos de seleção especialmente desenhados para SVMs testados, como RFE [30] e *Relief*-F [154]. Entretanto, o GAAcc selecionou subconjuntos de genes menores que o GAR2W2 e o GAJH. Por exemplo, para o conjunto de câncer de cólon os resultados obtidos foram: GAAcc (49 ± 30), GAR2W2 (382 ± 35) e o GAJH (388 ± 35).

Em PENG *et al.* [44], a seleção de genes informativos foi obtida pela utilização da RFE após a evolução por um GA. Diferente do trabalho de FRÖHLICH *et al.* [43], foi testado além do classificador SVM linear, outros empregando funções de núcleo polinomiais e Gaussianas. A função objetivo do GA era semelhante à apresentada por OOI e TAN [50], incluindo o erro de validação cruzada (LOOCV), onde o classificador MLHD foi substituído pelo classificador SVM. O algoritmo proposto foi aplicado aos dados de classificação binários de leucemia [17] e câncer de cólon [57], e aos dados de classificação multi-classes NCI60 [152] e GCM [39], com parâmetros evolutivos individualizados. Os melhores resultados (número de genes selecionados), foram obtidos pelo classificador SVM com núcleo polinomial de grau 4, com acurácia (erro de treinamento LOOCV) de 100% (6 genes), para o conjunto de leucemia; $93,55\%$ (12 genes), para o conjunto de câncer de cólon; $87,93\%$ (27 genes) para o conjunto de NCI60; e, $85,19\%$ (26 genes) para o conjunto de GCM. Poucos dos genes selecionados para o conjunto de leucemia concordavam com os publicados originalmente [17].

LIU *et al.* [45] usaram o algoritmo GA/SVM desenvolvido por PENG *et al.* [44], com tamanho populacional de 40 e cromossomo de tamanho 40 em 100000 gerações. Ao todo, 100 simulação foram realizadas com o conjunto de dados NCI60 [152], e acurácias de classificação LOOCV foram obtidas no intervalo entre $78,69\%$ e $88,52\%$. A contribuição dos autores encontra-se na caracterização dos genes selecionados, incluindo o uso de *nearest shrunken centroids* (NSC), análises de anotações e técnicas de mineração de texto na literatura, com as quais revelaram subclasses e marcadores tumorais.

SOUZA e CARVALHO [40] aplicaram uma algoritmo genético de tamanho populacional 100, 1000 gerações e imigração aleatória a cada 20 gerações de melhor solução única, para evitar convergência prematura, combinado com um classificador SVM multiclasses aos dados de Tumores de células pequenas, redondas e azuis (SRBCTs) [18]. Utilizando a validação cruzada aproximadora da generalização [56], ou GACV, como medida do erro de generalização em uma função objetivo semelhante a de FRÖHLICH *et al.* [43], encontrou-se uma acurácia média ($\pm$ desvio-padrão), em 10 simulações, de $0,980 \pm 0,258$, onde, em 6 delas, encontrou-se acurácia de 100%. Assim como em [43], a seleção do parâmetro de regularização foi incorporada no processo evolutivo para os valores de $C = \{0, 1, 1, 10, 100\}$.

HUERTA *et al.* [47] aplicaram um modelo de seleção de genes preditivos combinando lógica *fuzzy*, Algoritmos Genéticos e Máquina de Vetores de Suporte aos dados de leucemia [17] e câncer de cólon [57]. Os dados, assim como em [46], eram primeiro filtrados, eliminando aqueles de equivalência *fuzzy* determinada por cortes (α-*cuts*) em subconjuntos *fuzzy* Gaussianos. Com isso, os dados de leucemia foram reduzidos de 7129 para 1360 e os de câncer de cólon de 2000 para 943. O algoritmo GA/SVM empregado aos dados filtrados diferencia-se pela incorporação de uma etapa de coleção dos melhores genes após um número de gerações pré-determinado. Após o GA/SVM foi realizada uma análise de freqüência de seleção dos genes, levando a uma redução do número de genes para 50–100. Uma nova etapa do algoritmo GA/SVM é aplicado a esse sub-conjunto, determinando o classificador final. A acurácia média de classificação (número de genes), nas 5 simulações, dada pelo erro de LOOCV foi de 1 ± 0 ($32,8 \pm 8,228$ genes) e de $0,9929 \pm 0,9476$ ($12,5 \pm 2,8868$ genes), respectivamente para os dados de leucemia e câncer de cólon.

LIU e IBA [52], DEB e REDDY [53], DEB e REDDY [54] e PAUL e IBA [55] propuseram GAs de objetivos múltiplos (MOEA) como a solução para a seleção de genes em dados de microarranjos.

LIU e IBA [52] usaram um MOEA com três funções objetivo associado ao classificador de voto ponderado proposto por GOLUB *et al.* [17]. Os três objetivos eram: a razão de classificação incorreta do classificador; a diferença na razão de erro entre as classes; e, o tamanho do subconjunto de genes usados no classificador binário. As simulações foram executadas em três conjuntos de dados, referentes ao estudo de expressão gênica de leucemia [17], câncer de cólon [57] e linfoma [19], onde obtiveram acurácia média ($\pm$ desvio-padrão), em 10 simulações, respectivamente de $0,90 \pm 0,07$ ($15,2 \pm 4,54$ genes); $0,80 \pm 0,083$ ($11,4 \pm 4,27$ genes); e, $0,90 \pm 0,034$ ($12,9 \pm 4,40$ genes). Os melhores resultados de acurácia de classificação (n^o de genes) entre as 10 tentativas foram de 97% (16 genes), para os dados de leucemia; 90% (14 genes), para os dados de câncer de cólon; e, 94% (18 genes), para o conjunto de linfoma. As principais características incorporadas no MOEA foram a utilização da inicialização semi-aleatória de 10% dos alelos cromossômicos (genótipo) na representação binária da população inicial; a preservação da diversidade pela técnica de *Niche-Based Fitness Punishing* no espaço dos objetivos com distância de Hamming; e, seleção elitista com reinserção dos 10% melhores indivíduos entre gerações. O tamanho populacional empregado foi de 500 indivíduos e o critério de parada foi de 200000 avaliações da função objetivo (ou 400 gerações).

DEB e REDDY [53] usaram um GA de objetivos múltiplos (*non-dominated sorting EA* ou NSGA-II [155]) para minimizar o número de genes no classificador enquanto, simultaneamente, minimizava o número de amostras de treino e teste incorretamente classificadas. O NSGA-II apresenta as seguintes características: (1) utiliza a seleção elitista; (2) utiliza mecanismos de preservação da diversidade das soluções no espaço de objetivos, baseado em um valor de *crowding distance*; e (3) Enfatiza soluções não-dominadas, garantindo a convergência para soluções Pareto-ótimas. O classificador empregado na avaliação dos genes selecionados pelo NSGA-II foi o de voto ponderado [17]. O método foi aplicado aos dados de classificação binária de Leucemia [17], Linfoma difuso de grandes células B (DLBCL) [19], e câncer de cólon [57]. Os melhores valores de acurácia de classificação, dado pelo erro LOOCV na amostra de treinamento, foram de 100% para os três conjuntos de dados, com a utilização de 3 genes para os dados de leucemia, 5 genes para os dados de linfoma, e 6 genes para os dados de câncer de cólon. Em outro

trabalho, DEB e REDDY [54] ampliaram o emprego do NSGA-II com o classificador de voto ponderado à dados de classificação multi-classes. Os melhores resultados obtidos em simulações empregando os dados de GCM [39], com 14 classes, e, NCI60 [152], com 9 classes, foram, respectivamente, de 86% (erro de treinamento) e 80% (erro de teste), com 37 genes; e, $92,68\%$ (erro de treino) e 90% (erro de teste), com 12 genes.

Mais recentemente, PAUL e IBA [55] propuseram um novo método de computação evolutiva para a seleção de genes em dados de expressão. O método, intitulado *Model Building Genetic Algorithm* (PMBGA), que substitui o *crossover* e o operador de mutação tradicional pela construção de modelos probabilístico baseados nos dados (empírico) e por técnicas de amostragem para geração dos descendentes, foi associado aos classificadores empíricos de Bayes e o classificador do voto ponderado [17]. Simulações nos dados de leucemia [17], câncer de cólon [57] e linfoma [19] obtiveram acurácia média de classificação (50 tentativas), dada pela média entre o erro de validação cruzada *leave-one-out* do classificador *Naïve Bayes* na amostra de treinamento e o erro de teste ($n^{\underline{o}}$ de genes), respectivamente, de $0,96 \pm 0,03$ $(3,16 \pm 1)$; $0,96 \pm 0,02$ $(4,42 \pm 2,46)$; e, $0,88 \pm 0,04$ $(4,44 \pm 1,74)$, para os dados de leucemia, linfoma e câncer de cólon. Os mesmos resultados, quando da utilização do classificador do voto ponderado, foram de $0,94 \pm 0,03$ $(2,92 \pm 1)$; $0,94 \pm 0,02$ $(5,77 \pm 4,10)$; e, $0,83 \pm 0,04$ $(5,14 \pm 2,04)$.

Capítulo 4

Classificação de Perfis de Expressão Amostral por Máquinas de Vetores de Suporte

4.1 Máquinas de Vetores de Suporte (SVMs)

Máquinas de vetores de suporte são *sistemas de reconhecimento de padrões*, usualmente empregados em problemas de classificação binária, que definem funções discriminantes de margem ótima de separação entre classes.

Sistemas de reconhecimento de padrões desempenham classificações multi-classes, com múltiplos atributos, independente do tipo de regra de decisão aplicada. Um classificador atribui um padrão $\mathbf{x}$ à uma classe ω_k, $k \in \{1, 2, 3, \ldots, c\}$, particionando o espaço de atributos em segmentos lineares, áreas, volumes, e hiper-volumes, denominados regiões de decisão (fig. 4.1). A região de decisão (R_k) de uma classe pode ser descontínua, e as margens entre regiões de decisão adjacentes são denominadas margens de decisão, ou de separação.

Decisões de classificação baseadas em vetores de atributos $\mathbf{x}$ podem ser definidas pelo uso de funções discriminantes explicitamente definidas como,

$$d_k(\mathbf{x}), \qquad k = 1, 2, \ldots, c. \tag{4.1}$$

onde cada função discriminante está associada com uma classe particular conhecida ω_k, $k = 1, 2, \ldots, c$.

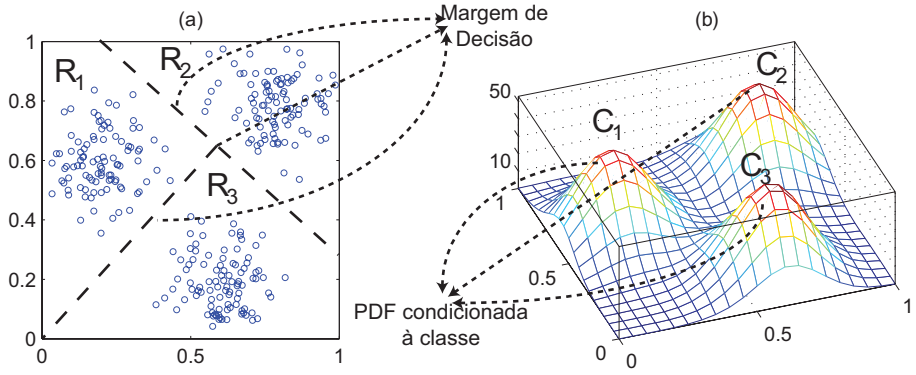

Figura 4.1: Regiões e Margens de decisão. (a) Regiões de decisão R_k e Margens de decisão para três classes não-sobrepostas; (b) Função de densidade de probabilidade associadas às classes

Em problemas dicotômicos, ou de classificação binária, emprega-se uma única função discriminante $d(\mathbf{x})$, que atribui classe de acordo com o sinal da resposta:

$$d(\mathbf{x}) = d_1(\mathbf{x}) - d_2(\mathbf{x}), \tag{4.2}$$

ou seja, quando $d_1(\mathbf{x})$ é maior que $d_2(\mathbf{x})$, $d(\mathbf{x}) > 0$, e o padrão $\mathbf{x}$ é atribuído à classe 1, caso contrário este é atribuído à classe 2.

Em problemas lineares dicotômicos, as tuplas de treinamento, dadas por $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_l}, y_l), \quad \mathbf{x} \in \mathbb{R}^n, \quad y \in \{-1, 1\}$, são linearmente separáveis, e há diferentes hiperplanos capazes de separar as classes. Usando os dados de treinamento durante a aprendizagem, o sistema de reconhecimento de padrões, expresso por uma máquina de aprendizado, encontra pesos $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ e o bias b, ou $1 \times w_0$, de uma função discriminante dada por:

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b, \quad \text{onde: } \mathbf{x}, \mathbf{w} \in \mathbb{R}^n. \tag{4.3}$$

Após treinamento bem sucedido, usando-se os parâmetros obtidos (eq. 4.3), a máquina de aprendizado, dado um padrão $\mathbf{x}$ desconhecido, produz uma saída f_o de acordo com a função indicadora dada por:

$$i_F = f_o = \operatorname{sgn}\left(d(\mathbf{x}, \mathbf{w}, b)\right). \tag{4.4}$$

ou:

$$\text{Se } d(\mathbf{x}, \mathbf{w}, b) > 0, \quad \mathbf{x} \in 1 \text{ (i.e., } o = y_1 = +1), \text{ e}$$
$$\text{Se } d(\mathbf{x}, \mathbf{w}, b) < 0, \quad \mathbf{x} \in 2 \text{ (i.e., } o = y_2 = -1).$$

A margem de decisão é então obtida pela intersecção entre a função discriminante (eq. 4.3), $d(\mathbf{x}, \mathbf{w}, b)$, e o espaço de atributos, sendo dada por,

$$d(\mathbf{x}, \mathbf{w}, b) = 0. \tag{4.5}$$

O aprendizado pode ser considerado como a descoberta da melhor função f_o de parâmetros ajustáveis $\mathbf{w}$ e b, utilizando os dados de treinamento disponíveis. Para se medir a qualidade de uma função qualquer f_o, deve-se definir medidas apropriadas (função de erro, custo ou perda), sendo as mais comuns: (1) Em regressão: O erro quadrático (norma L_2), definido como $L(y, f_o(\mathbf{x}, \mathbf{y})) = (y - f_o)^2$; e o erro absoluto (norma L_1), definido como $L(y, f_o(\mathbf{x}, \mathbf{y})) = |y - f_o|$; e (2) Em classificação (binária):

$$L(y, f_o(\mathbf{x}, \mathbf{y})) = 0, \quad \text{se } f_o(\mathbf{x}, \mathbf{y}) = y;$$
$$L(y, f_o(\mathbf{x}, \mathbf{y})) = 1 \quad \text{se } f_o(\mathbf{x}, \mathbf{y}) \neq y.$$

O Erro Médio, ou Risco Esperado, de uma função teórica de regressão $f(.)$, com função de erro dada pela norma L_2 é dada por:

$$R[f] = \mathbb{E}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 P(\mathbf{x}, y)\, D\mathbf{x}\, Dy, \tag{4.6}$$

onde, $(\mathbf{x}, y)$ são tuplas de treinamento e $f(\mathbf{x})$ é definida como a média de $P(y|\mathbf{x})$.

No aprendizado baseado em dados a função de densidade de probabilidade $P(\mathbf{x}, y)$ (eq. 4.6) é desconhecida, e o aprendizado deve basear-se nos dados amostrais (tuplas) de treinamento. O algoritmo de aprendizado de uma máquina de aprendizado deve descobrir a relação entrada-saída, função $f_o(\mathbf{x})$, usando apenas os dados de treinamento, o que, até certo ponto, aproxima-se da minimização do risco esperado $R[f]$ no espaço de busca T, ou seja,

$$f_o(\mathbf{x}) = arg \min_{f_o \in T} R[f]. \tag{4.7}$$

Uma maneira de proceder é usar os dados de treinamento para aproximar a integral no Risco Esperado (eq. 4.6) por uma soma finita. Isso leva à definição do Risco Empírico:

$$R_{emp}[f] = \frac{\sum_{i=1}^{l} (y_i - f(\mathbf{x}_i, \mathbf{w}))^2}{l}, \tag{4.8}$$

onde o conjunto de parâmetros $\mathbf{w}$ é o sujeito do aprendizado.

O *Princípio Indutor da Minimização do Risco Empírico* (ERM) recomenda que encontre-se a função $f(\mathbf{x}, \mathbf{w})$, que minimiza $R_{emp}[f]$ (eq. 4.8). A *Lei dos Grandes Números* garante que o Risco Empírico, $R_{emp}[f]$ (eq. 4.8), convirja para o Risco Esperado, $R[f]$ (eq. 4.6) quando o número de amostras de treinamento tende ao infinito ($l \rightarrow \infty$), ou seja,

$$\lim_{l \rightarrow \infty} (|R[f] - R_{emp}[f]|) = 0. \tag{4.9}$$

entretanto, não há garantias de que a função f_{emp}, que minimiza R_{emp}, convirja à melhor função f_o, que minimiza o Risco Esperado 4.7. O mesmo vale para os parâmetros $\mathbf{w}_{emp}$ e $\mathbf{w}_o$, que definem as funções f_{emp} e f_o, respectivamente.

O *Princípio da Minimização do Risco Estrutural* (SRM) é um princípio indutor de aprendizado em conjuntos finitos de dados de treinamento onde, a partir de um grande número de funções admissíveis ou máquinas de aprendizado com diferentes graus de liberdade, é escolhida uma função de complexidade correta (*capacidade*) para descrever a complexidade dos dados de treinamento (Minimização do Risco Estrutural). É muito útil no aprendizado a partir de dados amostrais pequenos, e tem como propósito do aprendizado combinar a complexidade dos dados de treinamento com a capacidade aproximadora (medida pela dimensão VC) do conjunto de funções que a máquina de aprendizado pode implementar.

A capacidade de um conjunto de funções indicadoras $i_F(\mathbf{x}, \mathbf{w})$ é expressa pela dimensão VC (fig. 4.2), que é o número máximo de pontos, h, que podem ser separados em todas as formas possíveis por essas funções. Por exemplo, a dimensão VC do hiperplano orientado em um espaço n-dimensional é igual a $n + 1$ (i.e., $h = n + 1$). Enquanto a capacidade de $\hat{f}_{n,1}$ aumenta (ex.: aumento do número de vetores de suporte), a capacidade de aproximação da máquina de aprendizado aumenta pelo uso de parâmetros ajustáveis adicionais, ou seja, menor grau de liberdade. Ao mesmo tempo, esse grande conjunto de parâmetros deve ser otimizado pelo uso da mesma quantidade de dados de treinamento, l, o que piora a estimativa do Risco Esperado. Ou seja, um aumento em n requer um aumento em l afim de garantir uma convergência uniforme entre os Riscos Empírico e Esperado.

A essência da SRM é a minimização de limites, que associam: (1) Erro de Generalização ($R(\mathbf{w}_n)$); (2) Erro de Aproximação ($R_{emp}(\mathbf{w}_n)$); (3) Dimensão VC (h); (4)

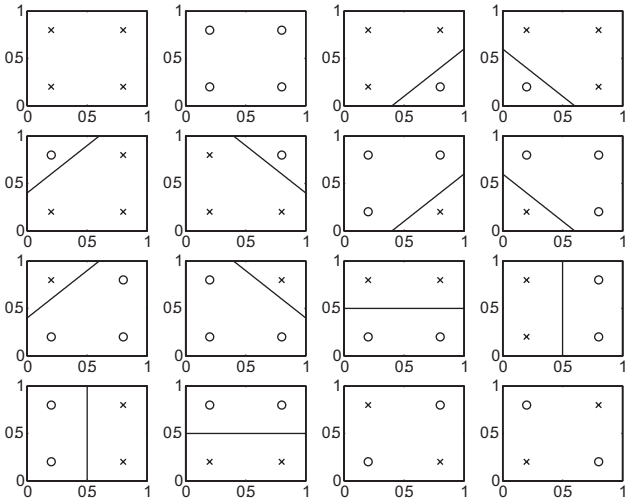

Figura 4.2: Dimensão VC para classificadores lineares. 4 pontos em $\mathbb{R}^{n=2}$ separadas em todas as formas $2^4 = 16$ possíveis pela função indicadora $i_F(\mathbf{x}, \mathbf{w}) = \text{sgn}\,(u)$, representada pela linha $u = 0$. Para $i_F(\mathbf{x}, \mathbf{w})$, $h = n + 1 = 3$. Os dois últimos casos inferiores correspondem ao problema clássico XOR, que não é separável por funções lineares.

Tamanho da amostra de treinamento (l); e, (5) Nível de confiança ($1 - \eta$). Na SRM o limite superior para o erro de generalização é dado por:

$$R(\mathbf{w}_n) \leq R_{emp}(\mathbf{w}_n) + \Omega\left(\frac{h}{l}, \frac{\ln \eta}{l}\right), \tag{4.10}$$

onde, o segundo termo do lado direito de (4.10) é chamado de *confiança VC* (termo de confiança ou intervalo de confiança) definido como

$$\Omega\left(\frac{h}{l}, \frac{\ln \eta}{l}\right) = \sqrt{\frac{h[(\ln(2l/h) + 1) - \ln(\eta/4)]}{l}}. \tag{4.11}$$

A notação $R(\mathbf{w}_n)$ indica que o risco é calculado no conjunto de funções aninhadas $f_n(\mathbf{x}, \mathbf{w}_n)$ de complexidade crescente, por exemplo, Máquinas de Vetor de Suporte de número de vetores de suporte crescente.

As equações (4.10 e 4.11) indicam que com número de dados de treinamento tendendo ao infinito ($l \to \infty$), o Risco Esperado $R(\mathbf{w}_n)$ é muito próximo ao Risco Empírico $R_{emp}(\mathbf{w}_n)$, porque $\Omega \to 0$. Por outro lado, quando a probabilidade $1 - \eta$ (nível de confiança) se aproxima de 1, o limite de generalização cresce consideravelmente, porque ao $\eta \to 0$ (ou $1 - \eta \to 1$), o valor de $\Omega \to \infty$ (fig. 4.3). Em outras palavras, qualquer máquina de aprendizado (modelo) obtido a partir de um número limitado de dados de

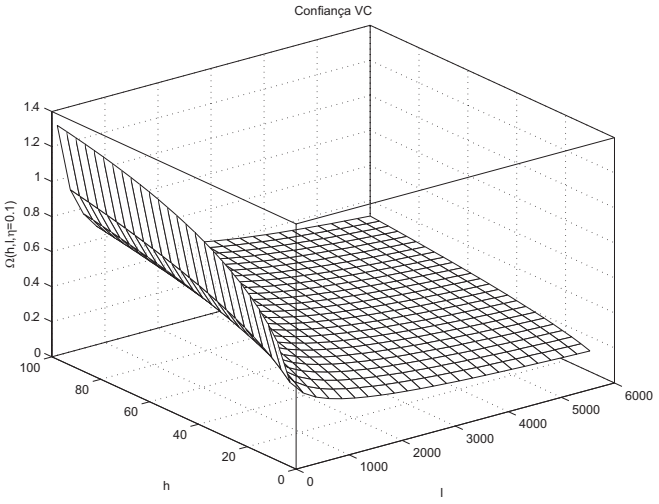

Figura 4.3: O intervalo de confiança VC. Dependência do intervalo de confiança VC, $\Omega(h, l, \eta)$, do número de dados de treinamento, l, e da dimensão VC, h, $h < l$, para um nível de confiança fixo $1 - \eta = 1 - 0,1 = 0,9$.

treinamento, não deve ter um intervalo de confiança VC arbitrário alto. Há sempre uma troca entre a acurácia fornecida pelo limite e o grau de confiança nesses limites.

A dimensão VC de SVMs pode ser extremamente alta ou infinita ($h = \infty$). Para manter o erro de generalização baixo, o intervalo de confiança é minimizado pela imposição de estrutura no conjunto de funções aproximadoras. Ou seja, para aplicar o SRM às SVMs, deve-se: (1) Introduzir estrutura em conjuntos aninhados S_A de hiperplanos canônicos com diferentes $\|\mathbf{w}\|$; e, (2) Escolher aquele com menor dimensão VC. Conjuntos S_A são analisados de modo que $\|\mathbf{w}\| \leq A$, onde $A_1 \leq A_2 \leq A_3 \leq \ldots \leq A_n$, resultando em conjuntos aninhados de funções $S_{A_1} \subseteq S_{A_2} \subseteq S_{A_3} \subseteq \ldots \subseteq S_{A_n}$ de capacidade e dimensão VC, h, crescentes.

Em $\mathbb{R}^n$ a distância entre um ponto $P(x_{1p}, x_{2p}, \ldots, x_{np})$ e um hiperplano $d(\mathbf{x}, \mathbf{w}, b) = 0$ (eq. 4.5), definido por $w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n \pm b = 0$, é dada por:

$$D = \frac{|(\mathbf{w}\mathbf{x}_p) \pm b|}{\|\mathbf{w}\|} = \frac{|w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n \pm b|}{\sqrt{w_1^2 + w_2^2 + w_3^2 + \ldots + w_n^2}}. \tag{4.12}$$

Impondo-se a restrição $\|\mathbf{w}\| \leq A$, o hiperplano canônico não pode estar mais próximo que $1/A$ de qualquer ponto de treinamento $\mathbf{x}_i$.

Vapnik [34] afirma que a dimensão VC, h, de um conjunto de hiperplanos canônicos em $\mathbb{R}^n$ restritos a $\|\mathbf{w}\| \leq A$ é dada por:

$$h \leq \min\left[R^2A^2, n\right] + 1, \tag{4.13}$$

onde todos os pontos de treinamento estão delimitados pela esfera de menor raio R (fig. 4.4). Um $\|\mathbf{w}\|$ pequeno resulta em um h pequeno, e a minimização de $\|\mathbf{w}\|$ é então a implementação do princípio da SRM.

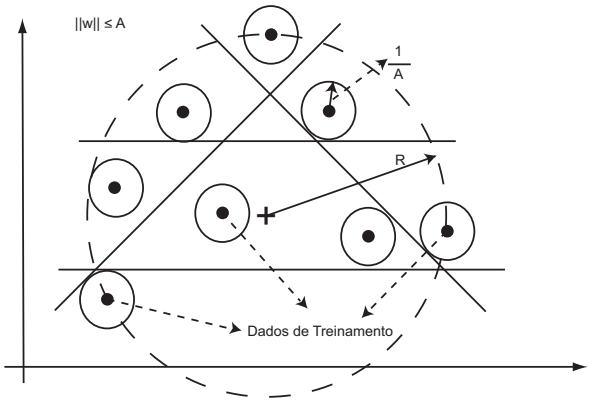

Figura 4.4: Aplicação do SRM à SVMs lineares. Restrição dos hiperplanos em permanecer fora das esferas de raio $1/A$ em torno de cada ponto dos dados de treinamento.

Com a disponibilidade apenas das tuplas de treinamento, além de encontrarmos o classificador que melhor aproxima o risco esperado pela minimização do risco empírico, desejamos encontrar entre todos os hiperplanos que minimizam o erro empírico, aquele de maior margem de separação, M, entre classes (fig. 4.5).

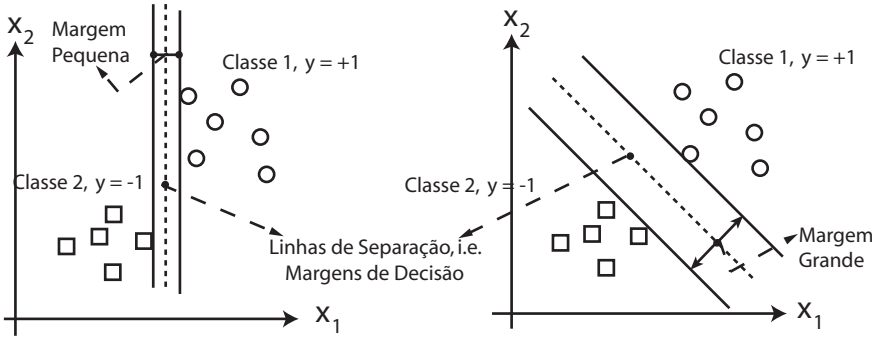

Figura 4.5: Planos de separação admissíveis. *Direita*, uma boa solução com margem grande, e *esquerda*, uma menos aceitável com margem pequena.

Geometricamente, a margem de separação M que deve ser maximizada durante o treinamento, é a projeção, no plano normal de pesos, da distância entre dois vetores de suporte quaisquer, pertencentes às duas diferentes classes. Essa margem é igual a:

$$M = (\mathbf{x}_1 - \mathbf{x}_2)_{\mathbf{w}} = (\mathbf{x}_1 - \mathbf{x}_3)_{\mathbf{w}}, \tag{4.14}$$

podendo ser encontrada como segue (fig. 4.5):

$$M = D_1 - D_2, \tag{4.15}$$

$$D_1 = \|\mathbf{x}_1\| \cos(\alpha) \quad \text{e} \quad D_2 = \|\mathbf{x}_2\| \cos(\beta), \tag{4.16}$$

$$\cos(\alpha) = \frac{\mathbf{x}_1^T \mathbf{w}}{\|\mathbf{x_1}\| \|\mathbf{w}\|} \quad \text{e} \quad \cos(\beta) = \frac{\mathbf{x}_2^T \mathbf{w}}{\|\mathbf{x_2}\| \|\mathbf{w}\|}, \tag{4.17}$$

$$\therefore M = \frac{\mathbf{x}_1^T \mathbf{w} - \mathbf{x}_2^T \mathbf{w}}{\|\mathbf{w}\|}. \tag{4.18}$$

Usando o fato de $\mathbf{x}_1$ e $\mathbf{x}_2$ serem vetores de suporte, i.e., $\mathbf{w}^T\mathbf{x}_1 + b = 1$ e $\mathbf{w}^T\mathbf{x}_2 + b = -1$, obtemos:

$$M = \frac{2}{\|\mathbf{w}\|}, \quad \text{onde: } \|\mathbf{w}\| = \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle} = \sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}. \tag{4.19}$$

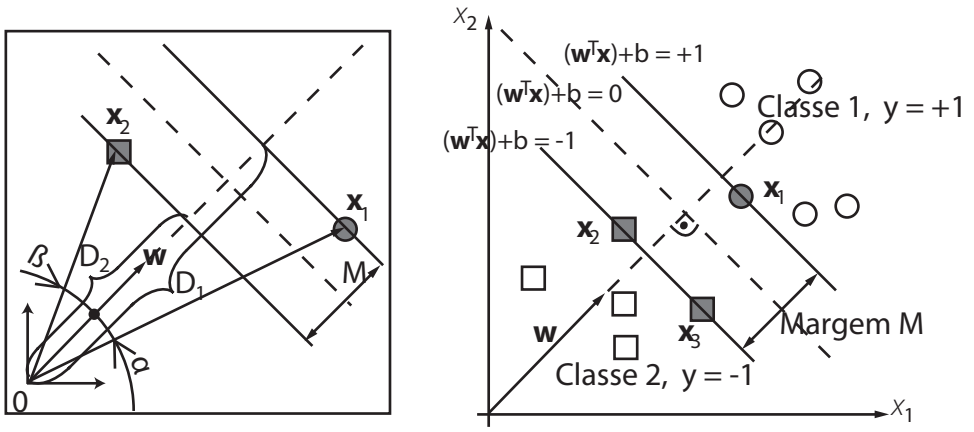

Figura 4.6: O hiperplano canônico de separação ótima (OCSH) e os Vetores de Suporte no espaço primordial. Os pontos satisfazendo $\mathbf{w}^T\mathbf{x}_1 + b = 1$ e $\mathbf{w}^T\mathbf{x}_2 + b = -1$ são vetores de suporte e o hiperplano canônico de separação ótima (OCSH), de maior margem, satisfaz $y_i|\mathbf{w}^T\mathbf{x}_i + b| \geq 1, i = 1, \ldots, l$.

A minimização da norma do vetor de peso $\|\mathbf{w}\|$ (eq. 4.19) é igual à minimização de $\mathbf{w}^T\mathbf{w} = \langle \mathbf{w}, \mathbf{w} \rangle = \sum_{i=1}^{n} w_i^2$, que leva à maximização da margem M. O *hiperplano canônico de separação ótima* (OCSH) definido pela margem $M = 2/\|\mathbf{w}\|$, especifica

vetores de suporte, que satisfazem $y_j|\mathbf{w}^T\mathbf{x}_j + b| = 1$, $j = 1, \ldots, N_{SV}$ (fig. 4.6). Ao mesmo tempo, o OCSH satisfaz as inequações:

$$y_i|\mathbf{w}^T\mathbf{x}_i + b| \geq 1, i = 1, \ldots, N_{SV}, \tag{4.20}$$

onde N_{SV} indica o número de vetores de suporte. Para encontrar o OCSH (fig. 4.7), uma máquina de aprendizado deve minimizar $\|\mathbf{w}\|^2$ sujeita à inequação 4.20. O problema de otimização não-linear com restrições em inequações é formulação no espaço primordial (espaço dos pesos), como:

$$\begin{aligned} &\min_{\mathbf{w},b} J_P(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w}, \\ &\quad \text{para: } y_i|\mathbf{w}^T\mathbf{x}_i + b| \geq 1, \quad i = 1, \ldots, l. \end{aligned} \tag{4.21}$$

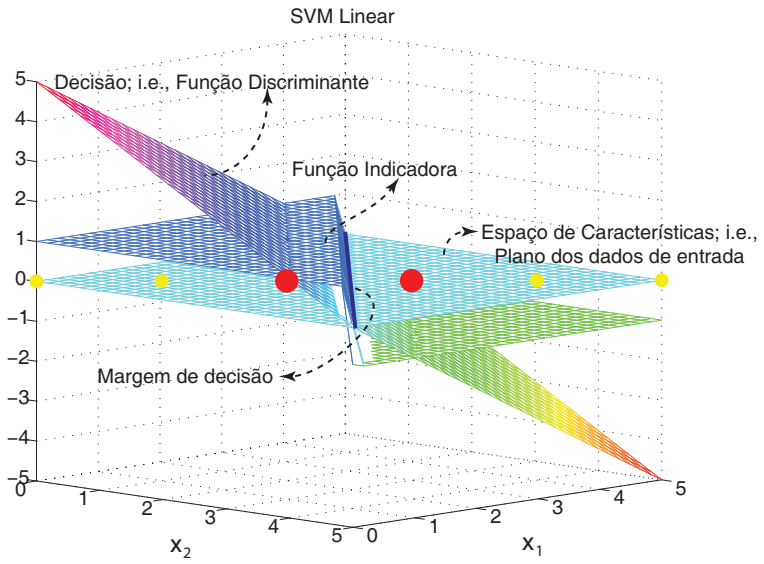

Figura 4.7: O hiperplano canônico de separação ótima (função de decisão), $d(\mathbf{x}, \mathbf{w}, b)$, margem de decisão (separação), $d(\mathbf{x}, \mathbf{w}, b) = 0$, e função discriminante (indicadora), $i_F = \text{sgn}\,(d(\mathbf{x}, \mathbf{w}, b))$.

A formulação no espaço dual (espaço dos multiplicadores de Lagrange) é formulada como:

$$\begin{aligned} &\mathcal{L}(\mathbf{w}; b; \alpha) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} - \textstyle\sum_{i=1}^{l} \alpha_i(y_i|\mathbf{w}^T\mathbf{x}_i + b| - 1), \\ &\quad \text{para: } \alpha_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \tag{4.22}$$

A solução é caracterizada pelo ponto estacionário da função de Lagrange (eq. 4.22):

$$\max_{\alpha} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b; \alpha) \tag{4.23}$$

Obtemos pela derivação parcial de $\mathcal{L}(.)$ que:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 & \rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i; \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0. \end{cases} \tag{4.24}$$

Pela expressão de $\mathbf{w}$ (obtida pela derivada parcial de $\mathcal{L}$ em função de $\mathbf{w}$) na função (eq. 4.24), transformamos a solução (eq. 4.23) em um problema de maximização em função dos multiplicadores de Lagrange α_i

$$\begin{aligned} &\max_{\alpha} J_D(\alpha) = \tfrac{1}{2} \sum_{i,j=1}^{l} y_i y_j \mathbf{x}_i^T \mathbf{x}_j \alpha_i \alpha_j + \sum_{i=1}^{l} \alpha_i, \\ &\text{para: } \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned} \tag{4.25}$$

gerando uma solução em função de $\alpha = [\alpha_1; \ldots; \alpha_l]$, e não mais de $\mathbf{w}$. A maximização na equação 4.25 resulta no classificador (fig. 4.8):

$$f(\mathbf{x}, \alpha, b) = \text{sgn} \,|\, \sum_{i=1}^{N_{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b|. \tag{4.26}$$

onde N_{SV} indica os vetores de suporte. Esse número de vetores de suporte, N_{SV}, pode ser bem menor que o número de dados de treinamento, l, gerando uma solução esparsa onde apenas os vetores de suporte são importantes em predições de classes para novos pontos.

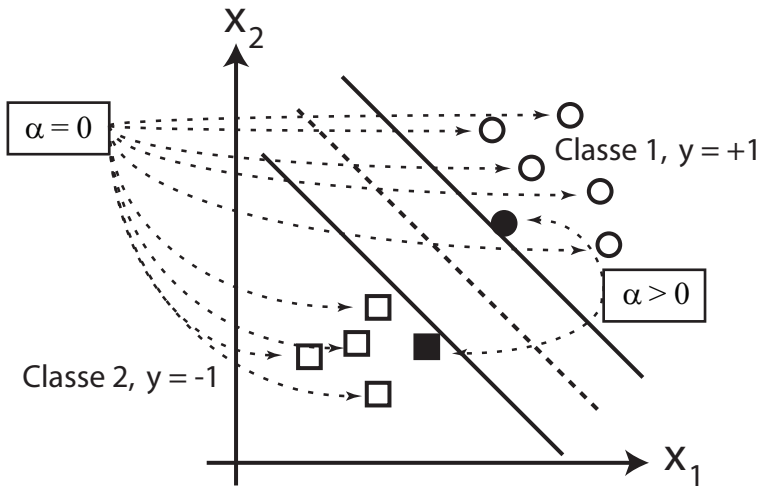

Figura 4.8: O hiperplano canônico de separação ótima (OCSH) e os Vetores de Suporte no espaço dual. Os pontos satisfazendo $\alpha = 0$ são vetores de suporte e o hiperplano canônico de separação ótima (OCSH), de maior margem é dado por $\text{sgn} \,|\, \sum_{i=1}^{N_{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b|$.

Encontrados os vetores de suporte, o limite superior no valor esperado de se cometer um erro de classificação em uma amostra independente de teste pode ser calculado como:

$$E_l[P(erro)] \leq \frac{E[N_{SV}]}{l}, \tag{4.27}$$

onde E_l indica o valor esperado sobre todos os l dados de treinamento. Este limite é independente da dimensão do espaço de entrada, ou seja, SVMs com pequeno número de vetores de suporte terão boa habilidade de generalização mesmo em espaços de entrada de alta-dimensão.

Em problemas reais é comum encontrarmos classes não-linearmente separáveis. Nesses casos, é possível construir classificadores lineares que tolerem classificações erradas pela inserção de variáveis escalares não negativas, $\xi_i (i = 1, \ldots, l)$, na formulação do problema [156]. O conjunto de inequações (eq. 4.20) é agora definido como:

$$y_i |\mathbf{w}^T \mathbf{x}_i + b| \geq 1 - \xi_i, i = 1, \ldots, l, \; \xi_i \geq 0, \tag{4.28}$$

ou,

$$\mathbf{w}^T \mathbf{x}_i + b \;\; \geq \;\; +1 - \xi_i, \quad \text{para } y_i = +1, \tag{4.29}$$

$$\mathbf{w}^T \mathbf{x}_i + b \;\; \geq \;\; -1 + \xi_i, \quad \text{para } y_i = -1. \tag{4.30}$$

Quando $\xi_i > 1$, a i-ésima inequação é violada quando comparada a inequação dos casos linearmente separáveis (eq. 4.20).

No espaço primordial o problema de otimização para o caso linear (eq. 4.21) se transforma em:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} J_P(\mathbf{w}, \xi) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi_i, \\ &\quad \text{para: } y_i |\mathbf{w}^T \mathbf{x}_i + b| \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \tag{4.31}$$

onde C é uma constante positiva e real. Obtemos então a seguinte função de Lagrange

$$\begin{aligned} &\mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu) = J_P(\mathbf{w}, \xi) - \sum_{i=1}^{l} \alpha_i (y_i |\mathbf{w}^T \mathbf{x}_i + b| - 1 + \xi_i) - \sum_{i=1}^{l} \nu_i \xi_i, \\ &\quad \text{para: } \alpha_i \geq 0, \quad \nu_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \tag{4.32}$$

Um segundo conjunto de multiplicadores de Lagrange ν_i é necessário pela inclusão das variáveis escalares ξ_i. A solução é dada pelo ponto estacionário da função de Lagrange:

$$\max_{\alpha,\nu} \min_{\mathbf{w},b,\xi} \mathcal{L}(\mathbf{w}, b, \xi; \alpha, \nu) \tag{4.33}$$

A minimização é resolvida pela derivação parcial de $\mathcal{L}$ em função das variáveis $\mathbf{w}$, b e ξ:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 & \rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i; \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0; \\ \frac{\partial \mathcal{L}}{\partial \xi_i} = 0 & \rightarrow \quad 0 \leq \alpha_i \leq C; \end{cases} \tag{4.34}$$

$$\text{para: } i = 1, \ldots, l. \tag{4.35}$$

Gerando o seguinte problema de otimização convexa:

$$\begin{aligned} &\max_{\alpha} J_D(\alpha) = -\tfrac{1}{2} \sum_{i,j=1}^{l} y_i y_j \mathbf{x}_i^T \mathbf{x}_j \alpha_i \alpha_j + \sum_{i=1}^{l} \alpha_i, \\ &\text{para: } \sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, l. \end{aligned} \tag{4.36}$$

Finalmente, a expressão para a função de decisão $d(\mathbf{x})$ do classificador SVM de margem *soft*, ou C–SVM (fig. 4.9), é a mesma do caso de classes linearmente separáveis (eq. 4.26), e conserva todas as propriedades das SVMs para classes linearmente separáveis.

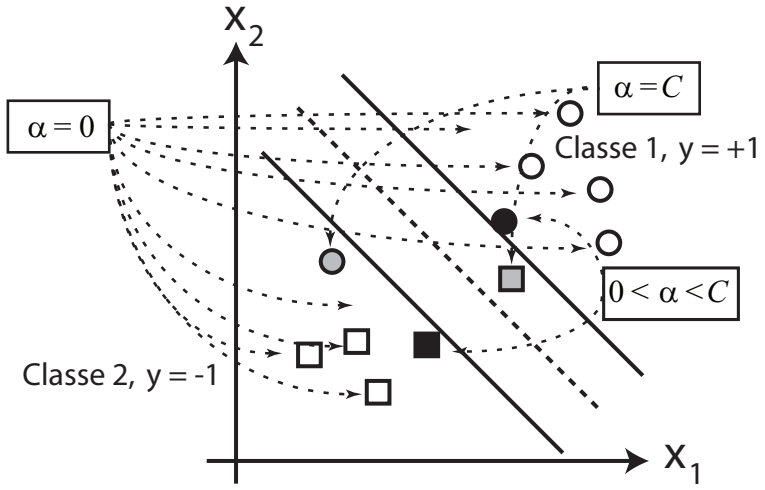

Figura 4.9: O hiperplano canônico de separação ótima (OCSH) e os Vetores de Suporte no espaço dual para problemas não-linearmente separáveis. Os pontos satisfazendo $0 < \alpha_i < C$ são vetores de suporte e o hiperplano canônico de separação ótima (OCSH), de maior margem é dado por $\text{sgn} \mid \sum_{i=1}^{N_{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b \mid$.

4.2 Classificação de Perfis de Expressão Amostral por Máquinas de Vetores de Suporte

Assim como na seleção de genes (seção 3.3), há duas formas de se pensar na classificação de dados de microarranjos [157]. BROWN *et al.* [158, 159] apresentaram as bases teóricas para a aplicação de SVMs na primeira, onde o questionamento experimental envolve a análise do padrão de expressão gênico. O propósito da análise é a descoberta de aglomerados/agrupamentos de genes correlacionados, que acredita-se, estão co-regulados na rede biomolecular. Métodos de classificação não-supervisionados, como a análise de aglomerados foram empregados nessa aplicação [160, 161, 162]. MUKHERJEE *et al.* [35], FUREY *et al.* [37] e GUYON *et al.* [30] apresentaram as bases teóricas para a aplicação de SVMs nas abordagem envolvendo a análise do padrão de expressão amostral, onde se conhece os rótulos de classes correspondentes a cada amostra. Outros métodos de classificação supervisionados foram usados com esse propósito [29, 28, 25].

BROWN *et al.* [158, 159] usaram SVMs para predizer funções de genes não caracterizadas do fungo *S. cerevisiae*, baseados em experimento de microarranjo de DNA em 79 amostras com 2467 genes de função conhecida. As SVMs foram empregadas no problema de classificação supervisionado de 6 classes funcionais de *S. cerevisiae* relacionadas ao ciclo do ácido tricarboxílico, ao ciclo respiratório, aos ribossomos citoplasmáticos, ao proteossoma, às histonas, e a proteínas do tipo *helix-turn-helix*; todas com pequeno número de exemplos positivos e muitos exemplos negativos. Os classificadores foram treinados com 3 amostras diferentes com $2/3$ dos dados, e o desempenho dos classificadores foi avaliado pelo erro de classificação médio com os $1/3$ restante dos dados. As SVMs lineares e com função de núcleo Gaussiana (RBF) e polinomial, foram comparadas a quatro outros métodos de classificação (*Parzen Windows*, Discriminante Linear de Fisher, e dois tipos de árvore de classificação (C4.5 e MOC1)) pelo erro médio de validação (LOOCV). O classificador de SVM treinado foi posteriormente empregado na predição de 3754 genes. Os autores relatam que o classificador SVM com função de núcleo dado por um função de base radial apresentou melhor resultado que os outros 4 classificadores avaliados para 5 classes, excetuando a classe das proteínas do tipo *helix-turn-helix*. Nenhum classificador avaliado foi capaz de reconhecer os genes dessa última classe.

MUKHERJEE *et al.* [35] foram quem primeiro propuseram o emprego de SVMs na análise do padrão de expressão gênico, ou ainda, na classificação de fenótipos celulares de tipos de câncer. Nesse trabalho, os classificadores de SVM foram utilizados nos dados de leucemia [17], onde a classe AML/ALL receberam rótulos ± 1. Além da classificação, a seleção de genes discriminantes foi abordada pela ordenação via correlação GS, onde além do conjunto de 7129 genes originais, foram selecionados mais três conjuntos de dados com os 999, 99 e 49 genes melhor ordenados, respectivamente. Para cada conjunto de dados, foram empregadas 38 amostras para treinamento e 34 para teste, seguindo a publicação original. Para os 4 conjuntos de dados, o erro de treinamento foi de zero amostras, e o erro de validação cruzada (LOOCV) foi, respectivamente, de 1, 0, 0 e 2, do conjunto de maior (7129) para o de menor (49) dimensão de entrada. Com a introdução de um nível de confiança na classificação, ou seja, pela eliminação das amostras classificadas com distância do hiperplano canônico de separação ótima (OCSH) menor que um distância pré-determinada $|d|$, a acurácia de treinamento (LOOCV) foi agora de zero amostras para todos os conjuntos, com amostras rejeitadas (nível de confiança), respectivamente, de $\approx 93\%$ (3), $\approx 95\%$ (0), $\approx 95\%$ (2), $\approx 93\%$ (4), para os conjuntos de maior para o de menor dimensão de entrada, com respectivos valores de $|d|$ de $0,1$, $0,08$, $0,08$ e $0,165$.

FUREY *et al.* [37] empregaram classificadores SVM lineares na classificação em dados de expressão de câncer de ovário, com 31 amostras de tecido canceroso, ovariano normal, e normal de outras origens, medidos em 97802 atributos. Conjuntos com dimensão de entrada de 25, 50, 100, 500 e 1000 atributos foram gerados pela ordenação via correlação GS. Os classificadores foram avaliados pelo erro de LOOCV. Apesar de não terem obtido bons resultados, foi possível corrigir problemas de amostras erroneamente rotuladas. A seleção dos atributos se mostrou de baixa relevância. Apenas 5 dos 10 atributos melhor ordenados eram genes funcionais, e destes, só 3 estavam relacionados com o câncer de ovário. Para avaliar a generalidade dos classificadores SVM, os dados de leucemia aguda [17] e câncer de cólon [57] foram também analisados e, os resultados comparados com o perceptron de camadas múltiplas. Para o primeiro conjunto, foram gerados dados com 25, 250, 500 e 1000 atributos melhor ordenados pela correlação GS, enquanto para o segundo foi gerado, além do conjunto original com 2000 atributos, outro com 1000 atributos melhor ordenados. Os resultados apresentados apontam para uma boa classificação para conjuntos maiores ou iguais a 1000 atributos, com acurácia de classificação semelhante aos trabalhos originais.

Como apresentado no capítulo anterior (seção 3.3), GUYON *et al.* [30] propuseram a seleção recursiva de genes discriminantes baseada nos pesos dos classificadores SVM treinados (SVM–RFE). Os resultados (número de genes selecionados), obtido pelo erro de validação cruzada (LOOCV), desse método nos conjuntos de dado de leucemia aguda [17] e câncer de cólon [57] foram, respectivamente de 100% (2) e 98% (4), indicando um grande benefício das SVM lineares de abordagens de ordenação recursiva de subconjuntos de genes. Diferente dos métodos de ordenação individual de genes adotados nos estudos anteriores, os 7 genes melhor ordenados pelo SVM–RFE estavam relacionados com os respectivos tipos de câncer.

Após esses trabalhos introdutórios da aplicação de SVMs na análise do perfis de expressão amostrais, trabalhos subseqüentes como os de BEN-DOR *et al.* [29] e POCHET *et al.* [163] ampliaram o número de conjuntos de dados de microarranjos de DNA analisados por classificadores SVM para problemas de duas classes, enquanto os trabalhos de YEANG *et al.* [164], RAMASWAMY *et al.* [39], MARKOWETZ *et al.* [165] e LEE *et al.* [166] sugeriram seu emprego também para problemas de classes múltiplas.

Capítulo 5

O Algoritmo de Seleção de Atributos em Microarranjos de DNA Proposto

O algoritmo proposto nesse estudo, consiste da seleção de perfis de expressão gênicos (atributos) via algoritmo genético de objetivos múltiplos (MOGA) combinado ao classificador binário de Máquinas de Vetores de Suporte de margem *soft* (C–SVM), aqui denominado SVM–MOGA, seguida da eliminação recursiva de atributos (SVM–RFE) dos perfis gênicos não preditivos selecionados no SVM–MOGA. O fluxograma da metodologia pode ser visto na figura 5.1, e suas etapas são descritas a seguir.

O SVM–MOGA é um algoritmo evolutivo que utiliza a abordagem de ordenação Pareto, onde o conceito de superioridade é explicitamente usado na determinação da probabilidade de seleção de cada indivíduo da população. A representação genotípica adotada para os indivíduos da população é a de cromossomos de dois genes (fig. 5.2) dados por vetores de valores binários $\{0, 1\}$. No primeiro gene, cada `bit` está associado com um perfil de expressão gênico, onde cada alelo desse vetor com valor igual a 1 indica que o perfil de expressão gênico participa do processo de classificação, enquanto o alelo com valor 0 indica o contrário. Além do primeiro gene, de comprimento igual ao número de perfis de expressão amostral dos conjuntos de dados pré-processados, cada cromossomo é acrescido de um segundo gene de dois `bits`, reservados à representação do parâmetro de regularização, C, da formulação da C–SVM (eq. 4.31), onde os genótipos $[0, 0]$, $[0, 1]$, $[1, 0]$ e $[1, 1]$ codificam, respectivamente, os fenótipos $\{0, 1, 10, 100, 1000\}$.

Dado o grande número de perfis gênicos disponíveis em estudos de expressão e a esperança de que poucos deles sejam realmente preditivos, a inicialização aleatória de

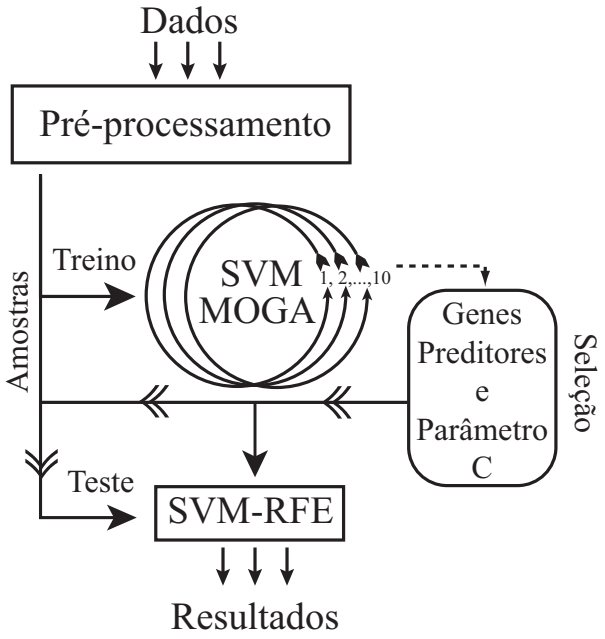

Figura 5.1: Fluxograma resumido da metodologia proposta. Os dados originais são processados para eliminação de perfis ruidosos, escalonamento e transformação logarítmica. As amostras de treinamento são normalizadas para distribuição normal padrão e as de teste escalonas de acordo. Os perfis gênicos são codificados para o MOGA, assim como parâmetros de regularização das C–SVM, e o processo de seleção ocorre por 10 simulações. Uma busca local por genes preditivos, entre os selecionados pelo SVM–MOGA, é obtida pela eliminação recursiva de atributos (RFE) com o parâmetro de regularização mais freqüente no SVM–MOGA, resultando na assinatura gênica final.

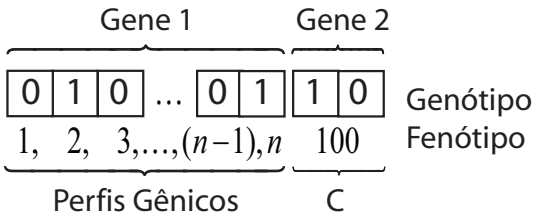

Figura 5.2: Representação genotípica adotada para o MOGA. Cada indivíduo (cromossomo) contém dois genes, o primeiro responsável pela codificação dos n atributos e o segundo responsável pela codificação do parâmetro de regularização C da C–SVM.

cada cromossomo é mantida apenas para os `bits` reservados ao segundo gene do MOGA. Para o primeiro gene do MOGA, uma inicialização semi-aleatória é escolhida, onde τ perfis gênicos são aleatoriamente selecionados e então codificados uniformemente com os valores 0 e 1. Há uma restrição de que ao menos τ_0 genes sejam selecionados. Ao final, são gerados n cromossomos, onde n é o tamanho populacional do MOGA, com no mínimo τ_0 e no máximo τ valores iguais a 1 no primeiro gene e, em média, um valor igual a 1 no segundo gene.

O SVM–MOGA contém três objetivos, $NObj = 3$, com funções f_1, f_2 e f_3, dadas, respectivamente, pela maximização da acurácia de classificação obtida pelo estimador *leave-one-out* de validação cruzada (LOOCV) com o classificador C–SVM; minimização da diferença absoluta da taxa de erro das classes de rótulo ± 1 do classificador C–SVM; e, minimização da razão dos perfis gênicos utilizados pelo classificador C–SVM. Apenas o subconjunto de treinamento é empregado no cálculo das funções objetivo. As funções f_1 e f_3 são empregadas para se obter o classificador C–SVM com melhor capacidade de generalização e de menor número de genes, enquanto a função f_2 é empregada para evitar-se *bias* de amostras não representativas das diferentes classes. Por exemplo, em um conjunto de dados com 98 perfis de expressão amostrais de rótulo de classe $+1$ e apenas 2 de uma segunda classe de rótulo -1, a função f_2 previne que o SVM–MOGA evolua para classificadores que predigam corretamente apenas a classe $+1$, obtendo acurácia de treino de 98%.

A acurácia de classificação, ou precisão total, de um classificador $f(.)$ é dada por $\frac{V_P+V_N}{n}$, onde V_P são as amostras de rótulo positivo $(+1)$ preditas como positivas, V_N são amostras de rótulo negativo (-1) preditas como negativas e n é o número total de amostras. Já a taxa de erro das classes ± 1 é dada por $\frac{F_N}{V_P+F_N}$ e $\frac{F_P}{F_P+V_N}$, respectivamente, para as classes de rótulo positivo e negativo, onde F_N são amostras de rótulo positivo preditas como negativas e F_P são amostras de rótulo negativo preditas como positivas. A relação dessas medidas para um problema de classificação binário são apresentadas na tabela 5.1.

O classificador SVM de margem *soft* (C–SVM) empregado tem função dada pela equação 4.26, repetida aqui por conveniência

$$f(\mathbf{x}, \alpha, b) = \operatorname{sgn} | \sum_{i=1}^{N_{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b|, \tag{5.1}$$

Tabela 5.1: Matriz de confusão para classificação binária.

Classe	**Predita** C_+	**Predita** C_-	**Taxa de Erro da Classe**
C_+	Verdadeiros positivos	Falsos negativos	$\frac{F_N}{V_P+F_N}$
	V_P	F_N	
C_-	Falsos positivos	Verdadeiros negativos	$\frac{F_P}{F_P+V_N}$
	F_P	V_N	

gerando uma solução esparsa, onde N_{SV} indica o número de vetores de suporte. O parâmetro de regularização C é empregado no problema de otimização convexa (eq. 4.36), permitindo a obtenção de uma margem *soft*.

Para cada experimento de microarranjo de DNA são realizadas 10 simulações, onde o SVM–MOGA evolui enquanto não é satisfeito um dos seguintes critérios de parada: (1) Número máximo de gerações; ou, (2) Valores máximo e mínimos para as funções objetivo experimento-dependentes. Não sendo satisfeito qualquer critério de parada, os indivíduos são modificados pelo operador de *crossover* de ponto único e mutação, com razões de *crossover* e de mutação selecionadas de forma empírica. Gerada a população intermediária pela modificação dos indivíduos da população atual, é atribuído a cada indivíduo o índice de aptidão baseado na ordenação Pareto ponderada pelo *fitness sharing*. A seleção é executada pelo operador *amostrador estocástico universal*.

Na manutenção da diversidade no processo evolutivo utiliza-se o *fitness sharing*, onde a distância genotípica, $d(.)$, entre dois cromossomos é dada pelas distância de Hamming e função de compartilhamento $sh(d)$ como na equação 3.7, repetida aqui por conveniência,

$$sh(d) = \begin{cases} 1 - (d/\sigma_{share})^{\alpha}, & \text{se } d < \sigma_{share}; \\ 0, & \text{caso contrário.} \end{cases} \tag{5.2}$$

onde α é uma constante que regula a forma da função de compartilhamento e, σ_{share} é o ponto de corte que regula a similaridade entre os cromossomos. A aptidão compar-

tilhada $F'(i)$ de um cromossomo i é dada pela equação 3.6, também repetida aqui por conveniência

$$F'(i) = \frac{F(i)}{\sum_{j=1}^{\mu} sh(d(i,j))}, \tag{5.3}$$

onde $F'(i)$ é igual à sua aptidão inicial, $F(i)$, dividida pela contagem no seu nicho. Essa função retorna 1 se os elementos são idênticos, 0 se são mais distantes que σ_{share}, e um valor nesse intervalo para níveis intermediários de dissimilaridade.

A re-inserção utiliza a estratégia elitista com a preservação de 10% dos indivíduos não-dominados acrescidos, quando necessário, dos melhores indivíduos selecionados pela ordenação Pareto em cada geração.

Ao final do SVM–MOGA, ou seja, quando o tempo máximo de gerações ou todos os critérios de parada é/são atendido(s), os perfis gênicos preditivos selecionados ao menos duas vezes nas soluções não-dominadas em cada uma das simulações são coletados em um subconjunto a ser reduzido pelo algoritmo SVM–RFE com classificador SVM de parâmetro de regularização C mais freqüentemente selecionado nesse mesmo subconjunto.

O algoritmo SVM–RFE (tab. 5.2) adotado segue o relato de GUYON *et al.* [30] onde, a cada iteração, os perfis gênicos são ordenados de acordo com seus pesos, $(w_i)^2$, e, aquele de menor peso quadrático é eliminado. O algoritmo é inicializado com todos os genes selecionados pelo SVM–MOGA e, os classificadores C–SVM obtido com os subconjuntos de genes selecionados são avaliados pela: (1) Acurácia de classificação de treinamento; (2) Acurácia de classificação de teste; e, (3) Acurácia de classificação total, ou seja, empregando-se tanto a amostra de treino como a de teste.

Tabela 5.2: Etapas do algoritmo SVM–RFE [30].

Entradas:	
$X_0 = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}^T$	% Dados de Entrada;
$\mathbf{y} = \{y_1, \ldots, y_N\}$	% Rótulos de classe;
Inicialização:	
$\mathbf{s} = [1, \ldots, N]$	% Subconjunto dos atributos selecionados;
$\mathbf{r} = [\quad]$	% Lista dos atributos selecionados;
Repetir até que $\mathbf{s} = [\quad]$:	
$X = X_0(:, \mathbf{s})$	% Restrição da amostra de treinamento aos atributos selecionados;
$\alpha = SVM(X, \mathbf{y})$	% Treinamento do classificador;
$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$	% Computação do vetor de pesos;
$c_i = (w_i)^2, \forall i$	% Computação do critério de ordenação;
$f = argmin(\mathbf{c})$	% Encontrar atributo de menor critério de ordenação;
$\mathbf{r} = [\mathbf{s}(f), \mathbf{r}]$	% Atualização da lista dos atributos selecionados;
$\mathbf{s} = \mathbf{s}(1 : f-1, f+1 : length(\mathbf{s}))$	% Eliminação do atributo de menor critério de ordenação;
Saída	
$\mathbf{r}$	% Lista de atributos ordenados.

Capítulo 6

Estudos Realizados

A seleção de atributos via RFE–SVM–MOGA proposta neste trabalho foi avaliada em três conjuntos de classificação binária empregando dados de microarranjos de DNA publicamente disponíveis. Os conjuntos de dados selecionados correspondem aos estudos de Leucemia Aguda [17], Linfoma difuso de grandes células B (DLBCL) [19] e Câncer de Cólon [57]. Um resumo das principais características desses conjuntos de dados pode ser visto na tabela 6.1.

Tabela 6.1: Descrição dos conjuntos de dados empregados no trabalho em função do número original de perfis gênicos (Número de Genes), diferentes rótulos de classe das amostras e tamanho das amostras (treino/teste) empregadas nos estudos realizados.

Conjunto de dados	Número de Genes	Classes	Tamanho das Amostras
Leucemia	7129	ALL/AML	47/25
Linfoma	4026	DLBCL/Outros	46/50
Cólon	2000	Normal/Câncer	22/40

Para os estudos apresentados a seguir, utilizou-se o *software* $MATLAB^{®}$ (The MathWorks, Inc.; http://www.mathworks.com/); o pacote $GEATbx$ [167], disponível em http://www.geatbx.com/, empregado na implementação do MOGA; assim como o pacote $LIBSVM$ [168], implementado em linguagem de programação C, e sua interface para $MATLAB^{®}$ desenvolvida pelos mesmos autores, empregados na otimização dos classificadores C–SVM.

6.1 Leucemia Aguda

O conjunto de dados de Leucemia aguda [17], disponível no sítio http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, contém o perfil de expressão amostral de 72 pacientes com leucemia aguda, oriundo de biópsias de medula óssea (62 casos) e sangue periférico (10 casos), medidos em 7129 sondas de 6187 genes humanos. Os dados estavam divididos em dois grupos: um conjunto de treinamento, com 27 amostras de leucemia linfoblástica aguda (ALL) e 11 amostras de leucemia mielóide aguda (AML), e um conjunto de teste independente, com 20 amostras de ALL e 14 amostras de AML.

O pré-processamento desses dados seguiu as etapas e parâmetros descritos em DEB e REDDY [53, 54] e PAUL e IBA [55], onde: (1) Os valores de intensidade negativos foram substituídos pela imposição de valores mínimos e máximos para 20 e 16000, respectivamente; (2) Analisando os perfis de expressão gênicos, excluíram-se aqueles com razão entre níveis de expressão máximo e mínimo ≤ 5 e diferença ≤ 500, restando 3859; e (3) Após a transformação logarítmica de base 10 dos valores de expressão, as amostras de treinamento e teste foram normalizadas pela subtração das médias dos perfis de expressão e divisão pelos desvios-padrão obtidos da amostra de treino. A transformação dos dados obtida pode ser vista nas figuras 6.1 e 6.2, respectivamente, para as amostras de treino e teste.

Na etapa de seleção de atributos via SVM–MOGA, além das funções de avaliação, f_1, f_2 e f_3, respectivamente, a acurácia de classificação obtida pelo estimador *leave-one-out* de validação cruzada, a diferença em módulo da taxa de erro das classes ± 1 e o número de perfis gênicos empregados no classificador, todas usando apenas a amostra de treinamento, o classificador teve seu desempenho estimado também pela acurácia de classificação de treino e teste. O parâmetro de razão de mutação adotado para esse conjunto de dados foi de $0,001$ $(4/3861)$, correspondendo a 4 `bits` alterados em média em cada cromossomo por geração. Os critérios de parada foram definidos como número máximo de gerações igual a 100; acurácia mínima de classificação de treinamento, f_1, de aproximadamente $0,95$ $(36/38)$, ou seja, tolerância de classificação errada de apenas duas amostras de treinamento; diferença absoluta entre a taxa de erro das classes de rótulo ± 1 do classificador C–SVM na amostra de treinamento, f_2, igual a zero; e, razão dos perfis gênicos utilizados pelo classificador C–SVM na amostra de treinamento, f_3, de aproximadamente $0,008$ $(30/3859)$, ou seja, seleção de 30 perfis gênicos entre os 3859

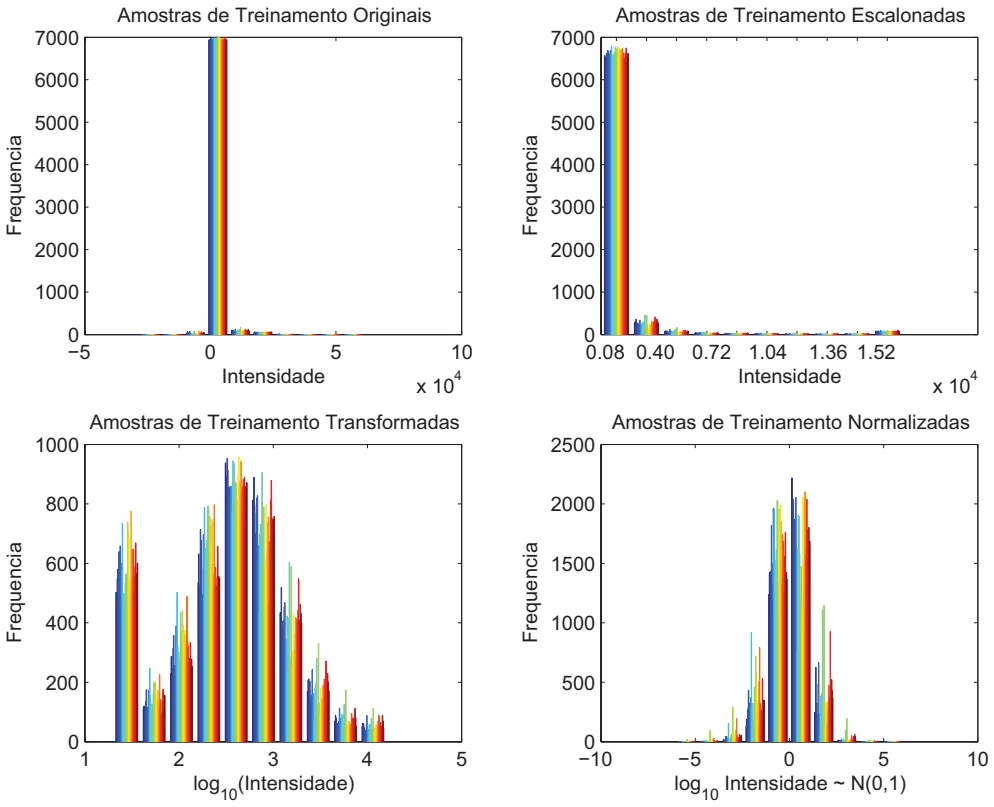

Figura 6.1: Pré-processamento das amostras de treinamento dos dados de Leucemia Aguda. Amostras de treinamento após a transformação dos dados de intensidade negativos e transformação logarítmica: antes da normalização (esquerda) e depois (direita).

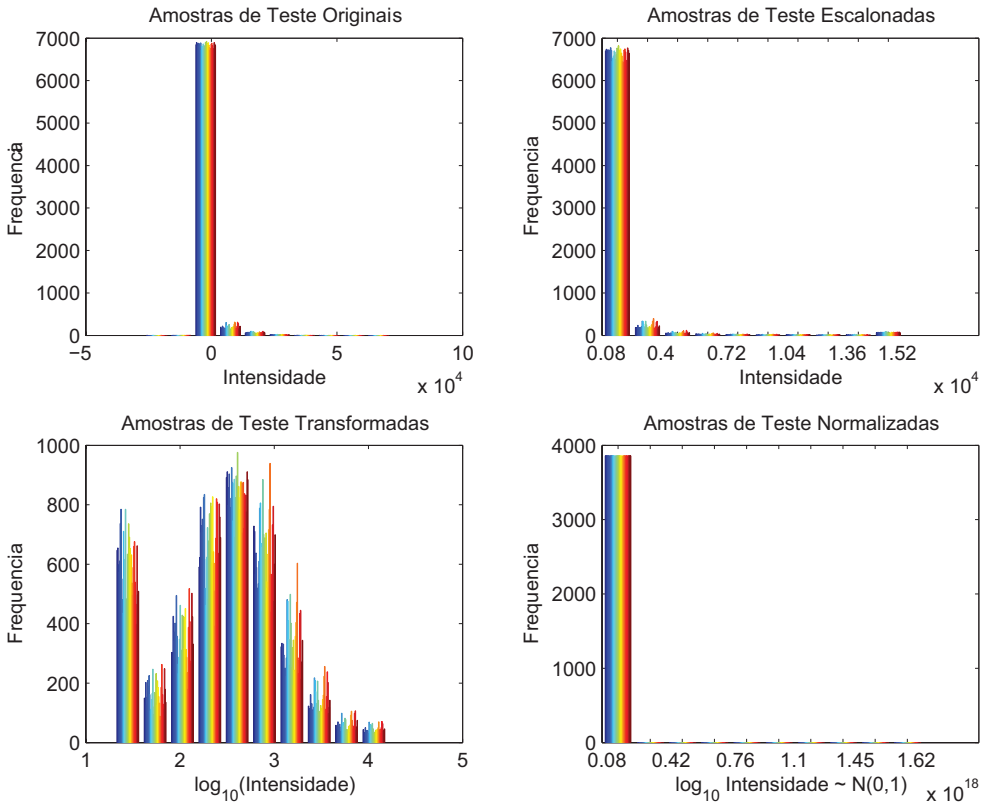

Figura 6.2: Pré-processamento das amostras de teste dos dados de Leucemia Aguda. Amostras de teste após a transformação dos dados de intensidade negativos e transformação logarítmica: antes da normalização (esquerda) e depois (direita).

resultantes da fase de pré-processamento. A relação dos demais parâmetros adotados no algoritmo SVM–MOGA estão resumidos na tabela 6.2.

Tabela 6.2: Parâmetros adotados nos estudos realizados para o algoritmo SVM–MOGA.

Simulações	10
População (n)	500
Gerações	100
Razão de *crossover*	$0,7$
Razão de Mutação	$\approx 0,001$
Razão de Re-inserção	0,9
τ_0	10
τ	100
α	1,0
σ_{share}	10,0

Os desempenhos dos classificadores com soluções pertencentes ao conjunto Pareto na geração final do SVM–MOGA, além do parâmetro de regularização C selecionado, são apresentados para cada uma das 10 simulações nas tabelas A.1 à A.10[1]. Um resumo desses resultados é apresentado na tabela 6.3, onde, para cada uma das 10 simulações, o melhor resultado está acompanhado da média ($\pm$ desvio-padrão) entre os classificadores do conjunto Pareto. Em média, o SVM–MOGA atendeu aos critérios de parada com 27 gerações ($\pm 14,85$), ou ainda com 12200 (± 6684) treinamentos de C–SVMs.

Dos 653 perfis gênicos presentes nos classificadores pertencentes aos conjuntos Pareto resultantes do SVM–MOGA nas 10 simulações, 161 tinham réplicas e foram incluídos na etapa de eliminação recursiva dos perfis gênicos não-preditivos (SVM–RFE). Na tabela 6.4 são apresentados os desempenhos dos classificadores C–SVM treinados com os 25 genes melhor ordenados pelo SVM–RFE com parâmetro de regularização $C = 10$, segundo a acurácia de classificação nas amostra de treinamento, teste e total (treinamento e teste).

A descrição e o símbolo fornecidos com os dados originais dos 15 genes cujos perfis de expressão foram empregados no classificador C–SVM de melhor desempenho são apresentados na tabela 6.5.

[1] Disponíveis na versão digital desse documento.

Tabela 6.3: Resumo dos resultados de treino com o conjunto de dados de Leucemia Aguda em função dos melhores valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas em cada uma das 10 simulações realizadas (média $\pm$ desvio-padrão).

Simulação	**Acurácia de Treino** (%)	**Acurácia LOOCV** (%)	**Dif. Razão de Erro**	**Acurácia de Teste** (%)	**C**	**N^{o} de Variáveis**
1	100	100	0	85,29	10	32
	(100 ± 0)	$(98,68 \pm 2,09)$	(0 ± 0)	$(83,82 \pm 2,94)$	(10 ± 0)	$(31,16 \pm 1,26)$
2	100	100	0	85,29	0,1	35
	(100 ± 0)	$(98,53 \pm 2,32)$	(0 ± 0)	$(79,73 \pm 4,75)$	$(34,6 \pm 49,15)$	$(33,44 \pm 2,35)$
3	100	97,36	0	61,76	100	30
	(100 ± 0)	$(98,24 \pm 1,51)$	(0 ± 0)	$(69,6 \pm 13,58)$	(100 ± 0)	$(33 \pm 5,19)$
4	100	94,73	0	79,41	10	28
	(100 ± 0)	$(96,49 \pm 3,03)$	(0 ± 0)	$(77,45 \pm 3,39)$	$(7 \pm 5,19)$	$(28,66 \pm 1,15)$
5	100	94,73	0	70,58	0,1	30
	(100 ± 0)	$(97,36 \pm 2,63)$	(0 ± 0)	$(78,23 \pm 7,66)$	$(0,64 \pm 0,49)$	$(31,4 \pm 1,51)$
6	100	100	0	88,23	10	29
	(100 ± 0)	$(99,12 \pm 1,51)$	(0 ± 0)	$(87,25 \pm 1,69)$	(10 ± 0)	$(28,66 \pm 0,57)$
7	100	94,73	0	73,52	10	30
	(100 ± 0)	$(96,49 \pm 3,03)$	(0 ± 0)	$(74,5 \pm 1,69)$	$(6,7 \pm 5,71)$	$(30,66 \pm 1,15)$
8	100	100	0	82,35	0,1	30
	(100 ± 0)	(100 ± 0)	(0 ± 0)	$(82,35 \pm 0)$	$(0,1 \pm 0)$	(30 ± 0)
9	100	97,36	0	76,47	100	30
	(100 ± 0)	$(98,68 \pm 2,09)$	(0 ± 0)	$(83,82 \pm 2,94)$	(10 ± 0)	$(31,16 \pm 1,26)$
10	100	100	0	70,58	10	30
	(100 ± 0)	$(98,94 \pm 2,35)$	(0 ± 0)	$(74,7 \pm 4,92)$	(10 ± 0)	$(29,2 \pm 1,78)$

Tabela 6.4: Resultado da etapa de eliminação recursiva de atributos para o conjunto de dados de Leucemia Aguda, segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} **de Atributos**	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	89,47	76,47	83,33
2	89,47	73,52	81,94
3	100	70,58	86,11
4	100	70,58	86,11
5	100	70,58	86,11
6	100	79,41	90,27
7	100	79,41	90,27
8	100	79,41	90,27
9	100	79,41	90,27
10	100	82,35	91,66
11	100	79,41	90,27
12	100	82,35	91,66
13	100	88,23	94,44
14	100	91,17	95,83
15	100	97,05	98,61
16	100	94,11	97,22
17	100	94,11	97,22
18	100	91,17	95,83
19	100	91,17	95,83
20	100	97,05	98,61
21	100	88,23	94,44
22	100	88,23	94,44
23	100	94,11	97,22
24	100	94,11	97,22
25	100	91,17	95,83

Tabela 6.5: Assinatura gênica do melhor classificador encontrado para o conjunto de dados de Leucemia Aguda.

Índice	**Descrição**	**Símbolo**
4499	CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7	X70297
1152	SPTB Spectrin, beta, erythrocytic	J05500
1926	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta	M31166
720	KIAA0260 gene, partial cds	D87449
6071	NF-kappa-B p65delta3 mRNA, spliced transcript lacking exons	U33838
2945	PTH2 parathyroid hormone receptor mRNA	U25128
5002	GB DEF = CD171 protein	Y10207
6847	Metallothionein I-B gene	M13485
2001	CSF1 Colony-stimulating factor 1 (M-CSF)	M37435
6806	Lysozyme gene (EC 3,2,1,17)	X14008
6279	PTX3 gene promotor region	X97748
4167	ALDR1 Aldehyde reductase 1 (low Km aldose reductase)	X15414
4095	FCER1A High affinity IgE receptor alpha-subunit (FcERI)	X06948
4823	Novel T-cell activation protein	X94232
2606	Actin bundling protein mRNA	U03057

6.2 Linfoma Difuso de Grandes Células B

O conjunto de dados de Linfoma difuso de grandes células B (DLBCL) [19], disponível no sítio http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt, contém 96 perfis de expressão gênicos, 54 normais e 42 de DLBCL, medidos usando microarranjos de cDNA contendo 4026 genes que são preferencialmente expressos em células linfóides, ou que tenham importância imunológica. Há, entretanto, em alguns arranjos, genes com valores de expressão indisponíveis. Para a correção dos valores de expressão ausentes, usou-se o algoritmo de k vizinhos próximos, proposto por TROYANSKAYA *et al.* [169], no qual k genes, com perfil de expressão similar ao gene com valor de expressão ausente, são selecionados e o valor de expressão desse gene é determinado pela média ponderada dos k-genes. Usamos o valor do parâmetro $k = 15$. Os dados foram aleatoriamente divididos em conjuntos de treinamento ($2/3$ dos dados, ou 63 perfis amostrais) e teste (1/3 dos dados, ou 33 perfis amostrais). Os valores de expressão das amostras de treino e teste foram normalizados para a distribuição normal padrão como no estudo anterior (seção 6.1).

Na etapa de seleção de atributos via MOGA combinada ao classificador C–SVM, assim como no estudo anterior, além das funções de avaliação, f_1, f_2 e f_3, cada classificador teve seu desempenho estimado também pela acurácia de classificação de treino e teste. Os desempenhos dos classificadores pertencentes ao conjunto Pareto na geração final do SVM–MOGA, além do parâmetro de regularização C selecionado, são apresentados para cada uma das 10 simulações nas tabelas A.13 à A.23[2]. Um resumo desses resultados é apresentado na tabela 6.6, onde, para cada uma das 10 simulações, o melhor resultado está acompanhado da média ($\pm$ desvio-padrão) entre os classificadores do conjunto Pareto. Foram adotados para o SVM–MOGA os mesmos parâmetros do estudo anterior 6.2 à exceção do parâmetro de razão de mutação, aqui de $0,001$ $(4/4028)$, correspondendo a 4 `bits` alterados em média em cada cromossomo por geração. Os critérios de parada foram definidos como número máximo de gerações igual a 100; acurácia mínima de classificação de treinamento, f_1, de aproximadamente $0,95$ $(60/63)$, ou seja, tolerância de classificação errada de apenas três amostras de treinamento; diferença absoluta entre a taxa de erro das classes de rótulo ± 1 do classificador C–SVM na amostra de treinamento, f_2, igual a zero; e, razão dos perfis gênicos utilizados pelo classificador C–SVM na amostra de treinamento, f_3, de aproximadamente $0,0075$ $(30/4026)$, ou seja, seleção de 30

[2] Disponíveis na versão digital desse documento.

perfis gênicos entre os 4026 existentes. Em média, o SVM–MOGA atendeu aos critérios de parada com $74,5$ gerações ($\pm 29,49$), ou ainda com 33575 ($\pm 1327,4$) treinamentos de C–SVMs.

Dos 643 perfis gênicos selecionados nos classificadores pertencentes aos conjuntos Pareto pelo SVM–MOGA nas 10 simulações, 296 tinham réplicas e foram incluídos na etapa de eliminação recursiva dos perfis gênicos não-preditivos (SVM–RFE). Na tabela 6.7 são apresentados os desempenhos dos classificadores C–SVM treinados com os 25 genes melhor ordenados pelo SVM–RFE com parâmetro de regularização $C = 1$, segundo a acurácia de classificação nas amostra de treinamento, teste e total (treinamento e teste).

A descrição e o símbolo fornecidos com os dados originais dos 16 genes cujos perfis de expressão foram empregados no classificador C–SVM de melhor desempenho são apresentados na tabela 6.8.

6.3 Câncer de Cólon

O conjunto de dados de Câncer de cólon [57], disponível em http://microarray.princeton.edu/oncology/affydata/index.html, contém 62 perfis de expressão gênicos de biópsias de cólon, 22 normais e 40 de câncer de cólon, medidos usando microarranjos de oligonucleotídeos de alta densidade contendo 2000 genes cada. Os dados foram aleatoriamente divididos em conjuntos de treinamento ($2/3$ dos dados, ou 40 perfis amostrais) e teste (1/3 dos dados, ou 22 perfis amostrais). Os perfis de expressão gênicos foram log-transformados (base 10) e as amostras de treino e teste normalizadas como nos estudos anteriores (seções 6.1 e 6.2).

Na etapa de seleção de atributos via MOGA combinada ao classificador C–SVM, além das funções de avaliação, f_1, f_2 e f_3, também como nos estudos anteriores, cada classificador teve ainda seu desempenho estimado pela acurácia de classificação de treino e teste. Os desempenhos dos classificadores pertencentes ao conjunto Pareto na geração final do SVM–MOGA, além do parâmetro de regularização C selecionado, são apresentados para cada uma das 10 simulações nas tabelas A.26 à A.35[3]. Um resumo desses resultados é apresentado na tabela 6.3, onde, para cada uma das 10 simulações, o melhor

[3] Disponíveis na versão digital desse documento.

Tabela 6.6: Resumo dos resultados de treino com o conjunto de dados de Linfoma Difuso de células B em função dos melhores valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBC e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas em cada uma das 10 simulações realizadas (média $\pm$ desvio-padrão).

Simulação	**Acurácia de Treino** (%)	**Acurácia LOOCV** (%)	**Dif. Razão de Erro**	**Acurácia de Teste** (%)	**C**	**N^{o} de Variáveis**
1	100	100	0	90,90	10	33
	(100 ± 0)	(100 ± 0)	(0 ± 0)	$(90,90 \pm 0)$	(10 ± 0)	(33 ± 0)
2	100	100	0	87,87	1	32
	(100 ± 0)	$(99,20 \pm 1,46)$	(0 ± 0)	$(89,77 \pm 1,56)$	$(25,75 \pm 45,82)$	$(30,75 \pm 2,31)$
3	100	95,23	0	93,93	1	29
	(100 ± 0)	$(97,61 \pm 2,18)$	(0 ± 0)	$(90,40 \pm 6,18)$	(1 ± 0)	$(32,83 \pm 3,71)$
4	98,41	95,23	0,03	87,87	0,1	28
	$(98,71 \pm 0,63)$	$(96,03 \pm 1,73)$	$(0,02 \pm 0,01)$	$(88,63 \pm 1,74)$	$(0,15 \pm 0,22)$	$(28,93 \pm 2,37)$
5	100	95,23	0	90,90	1	29
	(100 ± 0)	$(97,22 \pm 2,38)$	(0 ± 0)	$(90,90 \pm 2,47)$	(1 ± 0)	(34 ± 6)
6	100	100	0	90,90	1	32
	(100 ± 0)	$(99,08 \pm 1,52)$	(0 ± 0)	$(90,74 \pm 1,22)$	(1 ± 0)	$(31,36 \pm 1,11)$
7	98,41	96,82	0,03	90,90	0,1	31
	$(99,77 \pm 0,59)$	$(97,73 \pm 1,54)$	$(0 \pm 0,01)$	$(87,87 \pm 4,28)$	$(0,35 \pm 0,43)$	$(32,28 \pm 1,38)$
8	98,41	96,82	0,03	87,87	0,1	27
	$(99,29 \pm 0,83)$	$(97 \pm 1,47)$	$(0,01 \pm 0,01)$	$(89,22 \pm 3,07)$	$(0,4 \pm 0,45)$	$(28,55 \pm 1,94)$
9	100	100	0	90,90	100	33
	(100 ± 0)	$(99,94 \pm 0,28)$	(0 ± 0)	$(90,90 \pm 0)$	(100 ± 0)	$(32,86 \pm 0,73)$
10	100	98,41	0	90,90	10	30
	(100 ± 0)	$(98,94 \pm 0,91)$	(0 ± 0)	$(90,90 \pm 0)$	(10 ± 0)	$(31,33 \pm 2,30)$

Tabela 6.7: Resultado da etapa de eliminação recursiva de atributos para o conjunto de dados de Linfoma Difuso de células B, segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} de Atributos	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	52,38	51,51	52,08
2	52,38	51,51	52,08
3	68,25	60,60	65,62
4	66,66	51,51	61,45
5	73,01	66,66	70,83
6	76,19	69,69	73,95
7	85,71	84,84	85,41
8	85,71	87,87	86,45
9	93,65	84,84	90,62
10	95,23	78,78	89,58
11	96,82	72,72	88,54
12	96,82	81,81	91,66
13	98,41	72,72	89,58
14	98,41	63,63	86,45
15	98,41	72,72	89,58
16	100	90,9	96,87
17	100	90,9	96,87
18	100	90,9	96,87
19	100	90,9	96,87
20	100	90,9	96,87
21	100	87,87	95,83
22	100	87,87	95,83
23	100	87,87	95,83
24	100	87,87	95,83
25	100	87,87	95,83

Tabela 6.8: Descrição da assinatura gênica do melhor classificador encontrado para o conjunto de dados de Linfoma Difuso de células B.

Índice	Descrição	Símbolo
2384	Unknown; Clone=1186078	GENE1831X
573	Unknown UG Hs.136723 Homo sapiens HRIHFB2007 mRNA, partial cds; Clone=1283714	GENE3043X
1898	Similar to PCTAIRE2 kinase; Clone=1289872	GENE2652X
1240	Unknown UG Hs.135215 ESTs; Clone=1672325	GENE3300X
911	Unknown UG Hs.189061 ESTs; Clone=1318629	GENE2197X
600	IL-11 receptor alpha chain; Clone=1184280	GENE1936X
1808	Unknown UG Hs.193017 ESTs; Clone=1336500,	GENE2788X
14	Unknown UG Hs.136854 ESTs; Clone=1335756	GENE3117X
3862	Protocadherin 43 (PC43); Clone=704253	GENE3304X
2418	leukemia associated gene 1=13q14 gene deleted in CLL; Clone=299717	GENE933X
2644	Cell cycle progression 2 (CPR2); Clone=587981	GENE807X
304	Phospholipase C, gamma 2; Clone=1272756 (phosphatidylinositol-specific)	GENE3735X
3361	Unknown; Clone=1369321	GENE3846X
3365	Immunoglobulin alpha (1 or 2) heavy chain constant region; Clone=154441	GENE3850X
804	Unknown UG Hs.6179 Homo sapiens mRNA; Clone=1299811 cDNA DKFZp586K2322 (from clone DKFZp586K2322)	GENE2332X
3727	Similar to interferon-gamma inducible protein (MG11); Clone=1185239	GENE1596X

resultado está acompanhado da média ($\pm$ desvio padrão) entre os indivíduos do conjunto Pareto. Foram adotados para o SVM–MOGA os mesmos parâmetros dos estudos anteriores 6.2 à exceção do parâmetro de razão de mutação, aqui de $0,001$ $(2/2000)$, correspondendo a 2 `bits` alterados em média em cada cromossomo por geração. Os critérios de parada foram definidos como número máximo de gerações igual a 100; acurácia mínima de classificação de treinamento, f_1, igual a $0,95$ $(38/40)$, ou seja, tolerância de classificação errada de apenas duas amostras de treinamento; diferença absoluta entre a taxa de erro das classes de rótulo ± 1 do classificador C–SVM na amostra de treinamento, f_2, igual a zero; e, razão dos perfis gênicos utilizados pelo classificador C–SVM na amostra de treinamento, f_3, igual a $0,015$ $(30/2000)$, ou seja, seleção de 30 perfis gênicos entre os 2000 existentes. Em média, o SVM–MOGA atendeu aos critérios de parada com $9,8$ gerações $(\pm 1,81)$, ou ainda com 4460 $(\pm 816,08)$ treinamentos de C–SVMs.

Dos 503 perfis gênicos selecionados pelo SVM–MOGA nos classificadores pertencentes aos conjuntos Pareto nas 10 simulações, 122 tinham réplicas e foram incluídos na etapa de eliminação recursiva de atributos (SVM–RFE) dos perfis gênicos não-preditivos. Na tabela 6.10 são apresentados os desempenhos dos classificadores C–SVM treinados com os 25 genes melhor ordenados pelo SVM–RFE com parâmetro de regularização $C = 1$, segundo a acurácia de classificação nas amostra de treinamento, teste e total (treinamento e teste).

A descrição e o símbolo fornecidos com os dados originais dos 14 genes cujos perfis de expressão foram empregados no classificador C–SVM de melhor desempenho são apresentados na tabela 6.11.

Tabela 6.9: Resumo dos resultados de treino com o conjunto de dados de Câncer de Cólon em função dos melhores valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer de cólon, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas em cada uma das 10 simulações realizadas (média $\pm$ desvio-padrão). .

Simulação	**Acurácia de Treino** (%)	**Acurácia LOOCV** (%)	**Dif. Razão de Erro**	**Acurácia de Teste** (%)	**C**	**N^{o} de Variáveis**
1	100	97,5	0	68,18	100	28
	(100 ± 0)	$(98,33 \pm 1,44)$	(0 ± 0)	$(68,18 \pm 0)$	$(67 \pm 57,15)$	$(28,66 \pm 1,15)$
2	100	100	0	72,72	100	28
	(100 ± 0)	(100 ± 0)	(0 ± 0)	$(72,72 \pm 0)$	(100 ± 0)	(28 ± 0)
3	100	100	0	72,72	100	29
	(100 ± 0)	$(99,37 \pm 1,25)$	(0 ± 0)	$(70,45 \pm 2,62)$	$(50,5 \pm 57,15)$	$(28,5 \pm 1)$
4	100	100	0	77,27	10	33
	(100 ± 0)	$(99,16 \pm 1,29)$	(0 ± 0)	$(78,78 \pm 2,34)$	$(7 \pm 4,64)$	$(32 \pm 1,54)$
5	100	95	0	68,18	100	27
	(100 ± 0)	$(96,5 \pm 2,23)$	(0 ± 0)	$(70 \pm 4,06)$	$(40,06 \pm 54,71)$	$(28,6 \pm 2,3)$
6	100	100	0	68,18	10	25
	(100 ± 0)	(100 ± 0)	(0 ± 0)	$(68,18 \pm 0)$	(10 ± 0)	(25 ± 0)
7	100	100	0	72,72	1	30
	(100 ± 0)	(100 ± 0)	(0 ± 0)	$(72,72 \pm 0)$	(1 ± 0)	(30 ± 0)
8	100	97,5	0	77,27	1	30
	(100 ± 0)	$(97,5 \pm 1,58)$	(0 ± 0)	$(73,48 \pm 5,31)$	(1 ± 0)	$(29,83 \pm 0,98)$
9	100	97,5	0	77,27	0,1	25
	(100 ± 0)	$(98,33 \pm 1,44)$	(0 ± 0)	$(74,24 \pm 5,24)$	$(0,1 \pm 0)$	$(26,33 \pm 2,3)$
10	100	97,5	0	77,27	0,1	29
	$(99,37 \pm 1,25)$	$(97,5 \pm 2,04)$	$(0 \pm 0,01)$	$(75 \pm 4,54)$	$(0,1 \pm 0)$	$(29,25 \pm 1,25)$

Tabela 6.10: Resultado da etapa de eliminação recursiva de atributos para o conjunto de dados de Câncer de Cólon, segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} **de Atributos**	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	65	63,63	64,51
2	70	63,63	67,74
3	67,5	63,63	66,12
4	70	63,63	67,74
5	77,5	68,18	74,19
6	77,5	72,72	75,8
7	77,5	68,18	74,19
8	87,5	63,63	79,03
9	95	54,54	80,64
10	95	54,54	80,64
11	97,5	63,63	85,48
12	97,5	68,18	87,09
13	97,5	68,18	87,09
14	100	77,27	91,93
15	100	72,72	90,32
16	100	86,36	95,16
17	100	81,81	93,54
18	100	81,81	93,54
19	100	77,27	91,93
20	100	77,27	91,93
21	100	77,27	91,93
22	100	77,27	91,93
23	100	72,72	90,32
24	100	72,72	90,32
25	100	72,72	90,32

Tabela 6.11: Descrição da assinatura gênica do melhor classificador encontrado para o conjunto de dados de Câncer de Cólon.

Índice	**Descrição**	**Símbolo**
289	Retrovirus-related Env polyprotein (*Homo sapiens*)	R95874
1352	DNA-binding protein A (dbpA) gene, $3'$ end (*H. sapiens*).	M24069
1448	$214\,\mathrm{KD}$ Nucleoporin (*H. sapiens*)	H89481
1790	Heat Shock $27\,\mathrm{KD}$ protein (*H. sapiens*)	T48904
856	Heir-1 mRNA for helix-loop-helix protein (*H. sapiens*)	X66924
1913	DNA damage response protein kinase (DUN1) (*S. cerevisiae*)	R74066
499	Putative mucin core protein precursor 24 (*H. sapiens*); (contains L1 repetitive element)	D14043
14	Myosin light chain alkali, smooth-muscle isoform (*H. sapiens*)	H20709
1884	Mineralocorticoid receptor (*H. sapiens*)	R44301
801	ER lumen protein retaining receptor 1 (*H. sapiens*)	T47562
629	A42811 initiation factor (EIF-4A homolog)	T60318
205	Brain calcium channel BI-2 protein (*Oryctolagus cuniculus*)	H28452
452	Keratin 6 isoform K6e (KRT6E) mRNA, complete cds (*H. sapiens*)	L42611
698	Glia derived nexin precursor (*Mus musculus*)	T51261

Capítulo 7

Discussão

A seleção de perfis gênicos em microarranjos de DNA possibilita o aumento da acurácia de classificação, diminui o tempo de computação e, pela definição de uma assinatura com números reduzidos de genes, também possibilita o desenvolvimento de ferramentas de diagnóstico de menor custo, mais eficientes na pesquisa do câncer. Para classificação, além da seleção de genes altamente correlacionados com os rótulos de classe examinados, há ainda a necessidade de redução da redundância entre a informação contida nos mesmos. Como apresentado nesse trabalho, dados a complexidade e o tamanho do espaço de busca, métodos evolutivos de seleção de perfis gênicos associados à classificação supervisionada vem sendo empregados com esse propósito. Sendo a seleção de genes preditivos e o aumento da acurácia de classificação um problema intuitivamente multi-objetivo, algoritmos genéticos multi-objetivos ganharam grande destaque nesse propósito apresentando os melhores resultados entre os métodos de seleção envelopados.

Nesse trabalho apresentamos um novo método de seleção de perfis gênicos por algoritmo genético multi-objetivo (MOGA) associado ao classificador de Máquina de Vetores de Suporte (SVM) integrado à seleção recursiva de atributos. Foram apresentados estudos envolvendo a procura de marcadores tumorais em três conjuntos de dados binários de malignidades humanas em microarranjos de DNA.

A contribuição do trabalho consistiu no desenvolvimento de um MOGA para a seleção de perfis gênicos, de codificação genotípica especialmente desenhada para a codificação binária dos perfis gênicos e de constantes do classificador, de rápida convergência, de tamanho populacional reduzido e com critérios de parada explícitos. O emprego do conceito de dominância na seleção de conjuntos Pareto de genes candidatos aliada à eli-

minação recursiva de atributos nos permitiu obter classificadores paramétricos lineares de separação de margem, simples e de fácil interpretação, com boas acurácias de treinamento mesmo com número reduzido de perfis empregados na classificação e, com exceção do estudo em câncer de cólon, apresentando também boa capacidade de generalização.

Quando comparado aos achados de PAUL e IBA [55] nos mesmos conjuntos de dados, utilizando um algoritmo evolutivo de objetivos múltiplos (MOEA; *Multi-objective Evolutionary Algorithm*) semelhantes aos empregados aqui e originalmente proposto por LIU e IBA [52], associado a classificadores de voto ponderado (WV–MOEA), obtivemos (tab. 7.1) melhores resultados de treinamento empregando todos os dados (acurácia total) com número semelhante de perfis gênicos empregados no treinamento dos classificadores.

O método proposto nesta tese (RFE–SVM–MOGA) possui diversas vantagens em relação ao WV–MOEA. Há um menor custo computacional, observado pelo menor número máximo de treinamentos/avaliações do classificador—45050 em 100 gerações do SVM–MOGA contra 200000 em 500 gerações do WV–MOEA—bem como do uso inicial de menor número de perfis gênicos no treinamento de cada classificador—entre 10 e 100% no método proposto contra 10% no WV–MOEA. Além disso, há o emprego da informação contida nos melhores classificadores de cada uma das 10 simulações do algoritmo pelo emprego na eliminação recursiva dos genes presentes nas soluções pertencentes aos conjuntos Pareto na geração final. Em ambos trabalhos avaliando o método WV–MOEA [52, 55], os resultados são apresentados na forma de acurácia ou número de genes médios ($\pm$ desvio-padrão) das 10 simulações, não sendo estabelecido o critério de seleção do melhor classificador. Outra vantagem seria o emprego de um classificador linear, aqui SVMs de margem *soft*, ao invés de um classificador não-linear como o de voto ponderado [17]. O emprego de classificadores lineares, especificamente classificadores de margem de decisão, facilita a interpretação dos classificadores obtidos e do critério de decisão de classificação.

Outra comparação realizada foi a do método proposto com o emprego exclusivo da eliminação recursiva de atributos [30]. Os mesmas amostras de treino e teste pré-processados empregados no treinamento do RFE–SVM–MOGA tiveram a seleção de perfis gênicos realizada exclusivamente pela eliminação recursiva de atributos e os desempenhos dos classificadores C–SVM avaliados. Nas tabelas A.11, A.12, A.24, A.25, A.36, A.37 são apresentados, respectivamente, os desempenhos dos classificadores C–SVM treinados com os 25 genes melhor ordenados pelo SVM–RFE e a descrição e sím-

Tabela 7.1: Comparação dos resultados de RFE–SVM–MOGA com o método WV–MOEA em função da Acurácia Total (Acc. Total %) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos).

	Estudo	**Método**	
		RFE–SVM–MOGA	WV–MOEA
Acc. Total	Leucemia	$98,61$	90 ± 7
	Linfoma	$96,87$	90 ± 3
	Cólon	$91,93$	80 ± 8
Atributos	Leucemia	15	$15,2 \pm 4,54$
	Linfoma	16	$12,9 \pm 4,4$
	Cólon	14	$11,4 \pm 4,27$

bolos dos classificadores de melhor desempenho para os dados de leucemia, linfoma e câncer de cólon [1]. Na tabela 7.2 apresentamos essa comparação em função da Acurácia de Treinamento (Acc. Treino), Teste (Acc. Teste), Total (Acc. Total) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos). Para os dados de Leucemia, ambos os métodos apresentaram desempenho igual com 15 perfis gênicos (acurácia de treino, teste e total, respectivamente, de 100, $97,05$ e $98,61\%$), entretanto, o RFE–SVM obteve esse mesmo desempenho com a seleção de apenas 3, 4, 12, 13 e 14 genes. Assim como em trabalhos anteriores [82, 83], houve pouca sobreposição entre os 25 genes melhor ordenados pelos dois métodos, tendo em comum apenas o perfil do gene *X94232*, uma proteína ativadora de células T. Para os dados de Linfoma, o método proposto apresentou melhores resultados com relação às acurácias de teste ($90,90$ contra $87,87\%$) e total ($96,87$ contra $95,83\%$), bem como quanto ao menor número de perfis gênicos empregados (16 contra 25). A característica que diferencia esse conjunto de dados dos demais é a maior dimensão dos dados, tanto no que diz respeito ao número de amostras (96), quanto ao número de perfis gênicos após a fase de pré-processamento (4026). Finalmente, para os dados de câncer de cólon, o método proposto apresentou para o melhor classificador, com mesmo número de perfis gênicos, resultados inferiores aos encontrados pelo RFE–SVM. Enquanto com nosso método observamos acurácia de classificação de teste de apenas $77,27\%$, o emprego exclusivo do RFE–SVM obteve acurácia de teste de $86,36\%$. Como contraponto aos dados de Linfoma, esses são os dados

[1] Disponíveis na versão digital desse documento.

de menor dimensão entre os três estudados (68 amostras com 2000 genes). Não houve sobreposição entre os genes melhor ordenados pelos dois métodos nos dois casos estudados (dados de Linfoma e Câncer de Cólon). Os resultados obtidos aqui com o método RFE–SVM para os dados de Leucemia concordam com os apresentados por GUYON *et al.* [30], enquanto que para os dados de Câncer de Cólon são inferiores (Acc. de Teste de $86,36\%$ (14 genes) contra $90,32\%$ (8 genes)), entretanto esses últimos foram obtidos em amostras de validação cruzada LOOCV. Como discutiremos a seguir, resultados assim obtidos apresentam um *bias* de amostra que impossibilita a comparação justa entre métodos.

Tabela 7.2: Comparação dos resultados de RFE–SVM–MOGA com o método RFE–SVM em função da Acurácia de Treinamento (Acc. Treino $\%$), Teste (Acc. Teste $\%$), Total (Acc. Total $\%$) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos).

	Estudo	**Método**	
		RFE–SVM–MOGA	RFE–SVM
Acc. Treino	Leucemia	100	100
	Linfoma	100	100
	Cólon	100	100
Acc. Teste	Leucemia	$97,05$	$97,05$
	Linfoma	$90,9$	$87,87$
	Cólon	$77,27$	$86,36$
Acc. Total	Leucemia	$98,61$	$98,61$
	Linfoma	$96,87$	$95,83$
	Cólon	$91,93$	$95,16$
Atributos	Leucemia	15	3
	Linfoma	16	25
	Cólon	14	14

Para evitar a especialização de classificadores à uma amostra limitada de treinamento em dados de microarranjos de DNA, pesquisadores tem usado toda a amostra disponível e relatado a acurácia obtida durante a fase de seleção de genes como acurácia final; geralmente aquela obtida por técnicas de validação cruzada (k-*fold* ou LOOCV). Essa forma de estimação apresenta *bias* em relação aos dados disponíveis e pode ob-

ter predições pobres em amostra independente (amostra de teste). Ainda, nessa abordagem há a normalização dos perfis gênicos pela média e desvio-padrão obtidos a partir da amostra total, o que diminui a dificuldade de classificação de amostras independentes de diferente origem, como observado no caso dos dados de Leucemia Aguda (ver fig. 6.2). Exemplo usando esse enfoque pode ser visto em PENG *et al.* [44], onde foi apresentado um algoritmo similar ao desse trabalho, envolvendo a seleção evolutiva de perfis gênicos combinada a SVMs. Neste, um algoritmo genético adaptado de OOI e TAN [50] evoluiu 12 pequenos cromossomos de 36 a 40 `bits` durante 1219 e 31527 gerações, respectivamente, para os dados de Leucemia e Câncer de Cólon, associados a SVMs não-lineares com função de núcleo dada por polinômios de grau quatro que, a seguir, eram submetidos à eliminação recursiva de atributos [30]. Comparados aos resultados de acurácia total obtidos pelo método proposto (tab. 7.2), obtivemos $98,61\%$ (15 genes) e $91,93\%$ (14 genes) contra 100% (6 genes) e $93,55\%$ (12 genes) de acurácia LOOCV do trabalho de PENG *et al.* [44], respectivamente, para os dados de Leucemia e Câncer de Cólon. Mesmo com um classificador mais simples e linear, conseguimos resultados bastante similares sem o *bias* de amostra e com um algoritmo mais geral; ou seja, com mesmo número de gerações, tamanho populacional, codificação genotípica, entre outros, para os dois conjuntos de dados estudados. Especulamos que esse resultado seja conseqüência da seleção do parâmetro de regularização durante o processo evolutivo, prática não adotada em PENG *et al.* [44]. Outros exemplos onde a normalização e os resultados foram obtidos utilizando-se toda a amostra disponível incluem os trabalhos de FRÖHLICH *et al.* [43], JIRAPECH-UMPAI e AITKEN [46] e HUERTA *et al.* [47].

Outra dificuldade de comparação advém da inclusão da acurácia de teste como um dos objetivos da seleção evolutiva de perfis gênicos, como em DEB e REDDY [53, 54]. Apesar dos resultados obtidos nesses trabalhos serem os melhores que tomamos conhecimento, com a inclusão desse objetivo, a amostra de teste foi de fato empregada no treinamento do algoritmo de seleção e não pode ser considerada como independente.

Outros métodos evolutivos que permitem a comparação com o proposto nesse trabalho, exatamente por não terem *bias* de amostra, foram relatados por PAUL e IBA [55] tanto para o método por eles proposto, intitulado *Probabilistic Model Building Genetic Algorithm* (PMBGA), como pelo *Population Based Incremental Learning* (PBIL), proposto por BALUJA [170]. Ambos foram associados aos classificadores de voto ponderado (WV) e Naïve Bayes (NB) e aplicados aos 3 conjuntos de dados aqui também estudados.

Na tabela 7.3 apresentamos essa comparação em função da Acurácia de Treinamento (Acc. Treino), Teste (Acc. Teste), Total (Acc. Total) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos).

Tabela 7.3: Comparação dos resultados do método proposto (RFE–SVM–MOGA) com os métodos PMBGA e PBIL associados aos classificadores de voto ponderado (WV–PMBGA e WV–PBIL) e Naïve Bayes (NB–PMBGA e NB–PBIL) em função da Acurácia de Treinamento (Acc. Treino $\%$), Teste (Acc. Teste $\%$), Total (Acc. Total $\%$) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos).

	Estudo	**Método**				
		Proposto	WV–PMBGA	NB–PMBGA	WV–PBIL	NB–PBIL
Acc. Treino	Leucemia	100	100 ± 0	100 ± 0	100 ± 0	99 ± 1
	Linfoma	100	99 ± 1	99 ± 1	98 ± 2	99 ± 1
	Cólon	100	95 ± 3	95 ± 3	91 ± 4	91 ± 4
Acc. Teste	Leucemia	$97,05$	90 ± 6	90 ± 9	86 ± 6	80 ± 11
	Linfoma	$90,9$	93 ± 4	91 ± 4	91 ± 5	90 ± 6
	Cólon	$77,27$	81 ± 8	78 ± 8	77 ± 1	73 ± 9
Acc. Total	Leucemia	$98,61$	96 ± 3	95 ± 5	93 ± 3	90 ± 6
	Linfoma	$96,87$	96 ± 2	95 ± 2	94 ± 3	94 ± 3
	Cólon	$91,93$	88 ± 4	87 ± 4	84 ± 5	83 ± 5
Atributos	Leucemia	15	$3,16 \pm 1$	$2,92 \pm 1$	$10,8 \pm 7,14$	$10,2 \pm 7,99$
	Linfoma	16	$4,42 \pm 2,46$	$5,77 \pm 4,1$	$7,76 \pm 3,23$	$14,2 \pm 13,16$
	Cólon	14	$4,44 \pm 1,74$	$5,14 \pm 2,04$	$5,9 \pm 2,98$	$5,9 \pm 3,62$

Os resultados indicam que o método proposto obteve acurácia de treinamento ligeiramente melhor que os métodos evolutivos PMBGA e PBIL nos três estudos, sendo o único a obter acurácia de 100% com os dados de Câncer de Cólon. Quando avaliados pela acurácia de teste, nosso método apresentou resultado similares aos métodos PMBGA e PBIL, sendo ligeiramente superior ao método PBIL no estudo de Leucemia e ligeiramente inferior ao método PMBGA (WV–PMBGA e NB–PMBGA) no estudo de Câncer de Cólon. Quanto ao número de atributos (perfis de expressão) empregados no classificador final em cada um dos estudos, obtivemos resultados inferiores àqueles relatados com os métodos PBIL e PMBGA.

Os resultados experimentais relatados por diferentes autores e utilizados em comparações com aqueles encontrados pelo método proposto são resumidos na tabela 7.4. Em geral, em relação à capacidade de generalização (acurácia de teste), observamos melhores desempenhos do método não-evolutivo RFE–SVM para os dados de menor dimensão (dados de câncer de cólon) e dos métodos evolutivos (proposto, PBIL e PMBGA) para os dados de maior dimensão (dados de linfoma). À exceção do método PBIL e, possivelmente, do método MOEA (acurácia de teste não fornecida), todos os demais, evolutivos ou não, obtiveram boa capacidade de generalização com dados de dimensão média (Leucemia Aguda). Nos estudos realizados, observamos uma relação entre a acurácia de teste e o número de gerações do SVM–MOGA necessários à atender os critérios de parada. Enquanto para os dados de Linfoma, em 4 das 10 simulações a parada do SVM–MOGA ocorreu apenas pelo critério de gerações máximo (100 gerações), com média de $74,5$ ($\pm 29,49$) gerações, para os dados de Câncer de Cólon os critérios de parada foram atendidos prematuramente e o número médio de gerações para as 10 simulações foi de $9,8$ ($\pm 1,81$). Também observamos uma maior razão entre o número de perfis gênicos presentes nos classificadores pertencentes aos conjuntos Pareto com e sem réplicas nas 10 simulações nos dados de Linfoma, $2,17$ (296 com réplicas entre 643 perfis selecionados), em relação aos dados de Câncer de Cólon, $4,12$ (122 com réplicas entre 503 perfis selecionados), indicando a maior estabilidade do SVM–MOGA nos primeiros. Especulamos que os métodos evolutivos sejam mais apropriados a espaços de busca maiores, com maior número de perfis gênicos observados, e mais informativos, com maior número de amostras disponíveis, enquanto que métodos não-evolutivos são mais próprios para dados de menor dimensão. A ampliação de estudos em diferentes conjuntos de dados seria necessária à uma confirmação desses achados.

A ausência de informação dos melhores classificadores obtidos pelos métodos PMBGA, PBIL e MOEA e dos respectivos perfis gênicos selecionados impossibilitou qualquer comparação dessas observações. Apesar do bom desempenho dos métodos proposto e RFE–SVM para os dados de Leucemia, a confirmação dos achados em amostras independentes, ou mesmo nas amostras de teste utilizadas, por reação da cadeia de polimerase de tempo real (RT–PCR) não foi possível. Trabalhos desta natureza seriam necessários à comparação das assinaturas encontradas e elucidação da suficiência, ou insuficiência, de listas de genes tão pequenas quanto 3 à predição diferencial acurada em malignidades humanas.

Tabela 7.4: Comparação dos resultados de RFE–SVM–MOGA com os métodos WV–PMBGA, NB–PMBGA, WV–PBIL, NB–PBIL, WV–MOEA e RFE–SVM em função da Acurácia de Treinamento (Acc. Treino %), Teste (Acc. Teste %), Total (Acc. Total %) e pelo número de perfis gênicos empregados no treinamento dos classificadores (Atributos).

	Estudo	**Método**						
		RFE–SVM–MOGA	WV–PMBGA	NB–PMBGA	WV–PBIL	NB–PBIL	WV–MOEA	RFE–SVM
Acc. Treino	Leucemia	100	100 ± 0	100 ± 0	100 ± 0	99 ± 1	–	100
	Linfoma	100	99 ± 1	99 ± 1	98 ± 2	99 ± 1	–	100
	Cólon	100	95 ± 3	95 ± 3	91 ± 4	91 ± 4	–	100
Acc. Teste	Leucemia	$97,05$	90 ± 6	90 ± 9	86 ± 6	80 ± 11	–	$97,05$
	Linfoma	$90,9$	93 ± 4	91 ± 4	91 ± 5	90 ± 6	–	$87,87$
	Cólon	$77,27$	81 ± 8	78 ± 8	77 ± 1	73 ± 9	–	$86,36$
Acc. Total	Leucemia	$98,61$	96 ± 3	95 ± 5	93 ± 3	90 ± 6	90 ± 7	$98,61$
	Linfoma	$96,87$	96 ± 2	95 ± 2	94 ± 3	94 ± 3	90 ± 3	$95,83$
	Cólon	$91,93$	88 ± 4	87 ± 4	84 ± 5	83 ± 5	80 ± 8	$95,16$
Atributos	Leucemia	15	$3,16 \pm 1$	$2,92 \pm 1$	$10,8 \pm 7,14$	$10,2 \pm 7,99$	$15,2 \pm 4,54$	3
	Linfoma	16	$4,42 \pm 2,46$	$5,77 \pm 4,1$	$7,76 \pm 3,23$	$14,2 \pm 13,16$	$12,90 \pm 4,4$	25
	Cólon	14	$4,44 \pm 1,74$	$5,14 \pm 2,04$	$5,9 \pm 2,98$	$5,9 \pm 3,62$	$11,4 \pm 4,27$	14

Capítulo 8

Conclusão

Como apontado por YAO [171, 172], abordagens evolutivas são métodos de otimização que lidam facilmente com buscas em superfícies complexas, multi-modais e de grande dimensão. Entretanto, a maioria dos algoritmos evolutivos, segundo este, são ineficientes na busca local, podendo se beneficiar significativamente da incorporação de procedimentos de busca dessa natureza.

Nesse trabalho, tratamos o problema de seleção de grupos de perfis gênicos preditivos em dados de microarranjos de DNA, um problema de difícil solução exatamente pelas grandes dimensionalidade e complexidade do espaço de busca (espaço dos atributos) com um novo algoritmo de seleção, RFE–SVM–MOGA, que combina a busca paralela dos algoritmos genéticos com a busca local executado pela eliminação recursiva de atributos. Os resultados obtidos nos estudos realizados mostraram um bom desempenho, principalmente em dados de maiores complexidade e dimensão, com boa capacidade de generalização, obtidas sem *bias* de amostra. Quando comparado a outros algoritmos multi-objetivos, particularmente o MOEA [52], houve uma melhora do desempenho, com diminuição do custo computacional. O emprego exclusivo do método de eliminação recursiva, o RFE [30], mostrou bom desempenho, principalmente em espaços de busca menores, obtendo resultados melhores que os algoritmos multi-objetivos (MOGA e MOEA [52]), e mesmo que algoritmos evolutivos baseados em modelos probabilísticos (PMBGA [55] e PBIL [170]), em estudo realizado em dado com essa característica.

Apesar da necessidade de realização de um maior número de trabalhos comparando métodos evolutivos e recursivos em amostras sem *bias*, pode-se sugerir que métodos evolutivos tenham um melhor desempenho em dados de maior dimensão e com-

plexidade, cada vez mais disponíveis com a diminuição de custos do experimento, e que métodos recursivos, em geral mais rápidos, tenham um desempenho melhor em microarranjos menos densos, com espaço de perfis gênicos de menos dimensão.

Há a necessidade de comparação das assinaturas gênicas obtidas em amostras independentes e, ainda, por diferentes técnicas de biologia molecular (RT–PCR) para que se tenha real conhecimento da redução ideal do número de genes experimentados em microarranjos de DNA. A natural melhoria técnica da fabricação das lâminas de microarranjos de DNA e dos protocolos de purificação do mRNA, marcação e hibridização exigirá dos pesquisadores da área uma coleção de métodos de classificação e seleção de perfis gênicos para a obtenção de resultados mais fidedignos.

Nessa tese não foram estudados problemas multi-classes, onde amostras com k rótulos de classe são compradas simultaneamente em um único experimento. O emprego de heurísticas de combinação de classificadores SVM binários, como *um contra todos* ou *todos pareados*, permite a adaptação do método proposto para esse tipo de classificação [33, 34]. Espera-se um aumento da complexidade da seleção de grupos de genes preditivos pela necessidade de um maior número de margens de separação entre classes e relações entre as mesmas [173]. Uma dificuldade esperada nessa generalização do RFE–SVM–MOGA está relacionada ao cálculo, na amostra de treinamento, da capacidade de generalização, originalmente obtida pela validação cruzada *leave-one-out* (LOOCV). Propostas para a redução da complexidade computacional tida com cálculos repetidos da LOOCV nos $(k-1)$ classificadores necessários à separação das classes do problema são a utilização da estimativa do erro de validação cruzada aproximado [56], como em FRÖHLICH *et al.* [43] e SOUZA e CARVALHO [40], ou do *Bolstered error*, proposto por BRAGA-NETO e DOUGHERTY [174, 175].

Não foram realizadas comparações entre os resultados obtidos com o método proposto e outros obtidos por métodos de seleção individual de perfis gênicos. Métodos dessa última abordagem, especificamente aqueles envolvendo o cálculo da Informação Mútua entre os perfis gênicos e, desses com os rótulos de classe, foram investigados com sucesso em diversas aplicações, incluindo experimentos de microarranjos de DNA [176, 177, 178, 179]. Atualmente, um algoritmo de seleção individual de genes com cálculo da Informação Mútua obtida com a medida de entropia quadrática de Renyi [180] está sendo avaliado em dissertação de mestrado desenvolvida no Laborató-

rio de Engenharia de Sistemas de Saúde (LESS/PEB/UFRJ) e seu emprego hibridado ao método proposto, em uma etapa de pré-filtragem, poderá ser considerado.

Referências Bibliográficas

[1] WORLD HEALTH ORGANIZATION, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, *World Cancer Report*. Geneva, Switzerland, WHO Library, 2003.

[2] INSTITUTO NACIONAL DO CÂNCER, *Estimativa 2006: Incidência do Câncer no Brasil*. Brasil, 2005.

[3] JEGATHESAN, J., LIEBENTHAL, J., ARNETT, M., *et al.*, "Apoptosis: understanding the new molecular pathway", *Medsurg Nurs.*, v. 13, n. 6, pp. 371–375, Dec 2004. School of Nursing, University of Kansas, Kansas City, KS, USA.

[4] CAMPISI, J., "Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors", *Cell*, v. 120, n. 4, pp. 513–522, Feb 25 2005.

[5] LUCASSEN, A., WATSON, E., "Family history of breast cancer", *BMJ*, v. 330, n. 7481, pp. 26, Jan 1 2005.

[6] ECCLES, D., PICHERT, G., "Familial non-BRCA1/BRCA2-associated breast cancer", *Lancet Oncol.*, v. 6, n. 9, pp. 705–711, Sep 2005.

[7] LACROIX, M., LECLERCQ, G., "The "portrait" of hereditary breast cancer", *Breast Cancer Res Treat.*, v. 89, n. 3, pp. 297–304, Feb 2005.

[8] GARBER, J., OFFIT, K., "Hereditary cancer predisposition syndromes", *J Clin Oncol.*, v. 23, n. 2, pp. 276–292, Jan 2005.

[9] BODE, A., DONG, Z., "Signal transduction pathways in cancer development and as targets for cancer prevention", *Prog Nucleic Acid Res Mol Biol.*, v. 79, pp. 237–297, 2005.

[10] YU, E., HAHN, W., "The origin of cancer", *Cancer Treat Res.*, v. 122, pp. 1–22, 2004.

[11] STEELE, R., LANE, D., "P53 in cancer: a paradigm for modern management of cancer", *Surgeon*, v. 3, n. 3, pp. 197–205, Jun 2005.

[12] RISINGER, M., GRODEN, J., "Crosslinks and crosstalk: human cancer syndromes and DNA repair defects", *Cancer Cell*, v. 6, n. 6, pp. 539–545, Dec 2004.

[13] REARDON, J., SANCAR, A., "Nucleotide excision repair", *Prog Nucleic Acid Res Mol Biol.*, v. 79, pp. 183–235, 2005.

[14] ZHANG, Z., LI, M., RAYBURN, E., *et al.*, "Oncogenes as novel targets for cancer therapy (part I): growth factors and protein tyrosine kinases", *Am J Pharmacogenomics.*, v. 5, n. 3, pp. 173–190, 2005.

[15] ZHANG, Z., LI, M., RAYBURN, E., *et al.*, "Oncogenes as novel targets for cancer therapy (part II): Intermediate signaling molecules", *Am J Pharmacogenomics.*, v. 5, n. 5, pp. 327–338, 2005.

[16] WITTEKIND, C., NEID, M., "Cancer invasion and metastasis", *Oncology*, v. 69, n. Suppl 1, pp. 14–16, Sep 19 2005.

[17] GOLUB, T. R., SLONIM, D. K., TAMAYO, P., *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.", *Science*, v. 286, n. 5439, pp. 531–537, October 1999.

[18] KHAN, J., WEI, J. S., RINGNER, M., *et al.*, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks", *Nat. Med.*, v. 7, pp. 673–679, 2001.

[19] ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.", *Nature*, v. 403, n. 6769, pp. 503–511, Feb 2000.

[20] MA, X. J., WANG, Z., RYAN, P. D., *et al.*, "A Two-Gene Expression Ratio Predicts Clinical Outcome in Breast Cancer Patients Treated with Tamoxifen", *Cancer Cell*, v. 5, pp. 607–616, 2004.

[21] CHEN, C. D., WELSBIE, D. S., TRAN, C., *et al.*, "Molecular Determinants of Resistance to Antiandrogen Therapy", *Nat. Med.*, v. 10, pp. 33–39, 2004.

[22] SRINIVAS, P. R., KRAMER, B. S., SRIVASTAVA, S., "Trends in Biomarker Research for Cancer Detection", *Lancet Oncol.*, v. 2, pp. 689–704, 2001.

[23] SCHUCHHARDT, J., BEULE, D., MALIK, A., *et al.*, "Normalization Strategies for cDNA Microarrays", *Nucleic Acids Res*, v. 28, pp. E47, 2000.

[24] KOHAVI, R., JOHN, G. H., "Wrappers for Feature Subset Selection.", *Artif. Intell.*, v. 97, n. 1-2, pp. 273–324, 1997.

[25] DUDOIT, S., FRIDLYAND, J., SPEED, T. P., "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", *Journal of the American Statistical Association*, v. 97, n. 457, pp. 77–87, 2002.

[26] GOH, L., SONG, Q., KASABOV, N., "A novel feature selection method to improve classification of gene expression data". In: *CRPIT '29: Proceedings of the second conference on Asia-Pacific bioinformatics*, pp. 161–166, Darlinghurst, Australia, 2004.

[27] PARK, P. J., PAGANO, M., BONETTI, M., "A nonparametric scoring algorithm for identifying informative genes from microarray data.", *Pac Symp Biocomput*, pp. 52–63, 2001.

[28] KELLER, A., SCHUMMER, M., HOOD, L., *et al.*, *Bayesian Classification of DNA Array Expression Data*, Report -2000-08-01, UW-CSE, 2000.

[29] BEN-DOR, A., BRUHN, L., FRIEDMAN, N., *et al.*, "Tissue classification with gene expression profiles.", *J Comput Biol*, v. 7, n. 3-4, pp. 559–583, 2000.

[30] GUYON, I., WESTON, J., BARNHILL, S., *et al.*, "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, v. 46, n. 1-3, pp. 389–422, 2002.

[31] FURLANELLO, C., SERAFINI, M., MERLER, S., *et al.*, "Entropy-based gene ranking without selection bias for the predictive classification of microarray data.", *BMC Bioinformatics*, v. 4, pp. 54, Nov 2003.

[32] FURLANELLO, C., SERAFINI, M., MERLER, S., *et al.*, "An accelerated procedure for recursive feature ranking on microarray data.", *Neural Netw*, v. 16, n. 5-6, pp. 641–648, 2003.

[33] VAPNIK, V. N., *The Nature of Statistical Learning Theory*. New York, USA, Springer-Verlag, 1995.

[34] VAPNIK, V. N., *Statistical Learning Theory*. New York, USA, John Wiley and Sons, Inc., 1998.

[35] MUKHERJEE, S., RIFKIN, R., *Support Vector Machine Classification of Microarray Data*, A. I. Memo 1677, Artificial Intelligence Laboratory and The Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, 1999.

[36] MUKHERJEE, S., TAMAYO, P., SLONIM, D., *et al.*, *Support Vector Machine Classification of Microarray Data*, A.I. Memo 1677, Massachusetts Institute of Technology Artificial Inteligence Laboratory and Center for Biological and Computational Learning Department of Brain and Cognitive Sciences, 2000.

[37] FUREY, T. S., CRISTIANINI, N., DUFFY, N., *et al.*, "Support vector machine classification and validation of cancer tissue samples using microarray expression data.", *Bioinformatics*, v. 16, n. 10, pp. 906–914, Oct 2000.

[38] FUREY, T. S., CRISTIANINI, N., DUFFY, N., *et al.*, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data", *Bioinformatics*, v. 16, n. 10 2000, pp. 906–914, 2000.

[39] RAMASWAMY, S., TAMAYO, P., RIFKIN, R., *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures.", *Proc Natl Acad Sci U S A*, v. 98, n. 26, pp. 15149–15154, Dec 2001.

[40] SOUZA, B. F. D., CARVALHO, A. P. D. L. F. D., "Gene Selection Based on Multi-Class Support Vector Machines and Genetic Algorithms", *Genetics and Molecular Research*, v. 4, n. 3, pp. 599–607, 2005.

[41] HOLLAND, J. H., *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA, University of Michigan Press, 1975.

[42] GOLDBERG, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA, Addison-Wesley, 1989.

[43] FROHLICH, H., CHAPELLE, O., SCHOLKOPF, B., "Feature Selection for Support Vector Machines by Means of Genetic Algorithms", *ICTAI*, v. 0, pp. 142, 2003.

[44] PENG, S., XU, Q., LING, X. B., *et al.*, "Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines", *FEBS Lett*, v. 555, pp. 358–362, 2003.

[45] LIU, J. J., CUTLER, G., LI, W., *et al.*, "Multiclass cancer classification and biomarker discovery using GA-based algorithms.", *Bioinformatics*, v. 21, n. 11, pp. 2691–2697, Jun 2005.

[46] JIRAPECH-UMPAI, T., AITKEN, S., "Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes.", *BMC Bioinformatics*, v. 6, pp. 148, 2005.

[47] HUERTA, E. B., DUVAL, B., HAO, J.-K., "A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data". In: *EvoWorkshops*, pp. 34–44, 2006.

[48] LI, L., WEINBERG, C. R., DARDEN, T. A., *et al.*, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/kNN method", *Bioinformatics*, v. 17, n. 12, pp. 1131–1142, 2001.

[49] LIU, J., HITOSHI, I., "Selecting Informative Genes with Parallel Genetic Algorithms in Tissue Classification", *Genome Informatics*, v. 12, pp. 14–23, 2001.

[50] OOI, C. H., TAN, P., "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data", *Bioinformatics*, v. 19, n. 1, pp. 37–44, 2003.

[51] LIN, T., LIU, R., CHEN, S., *et al.*, "Genetic Algorithms and Silhouette Measures Applied to Microarray Data Classification." In: *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, 2005.

[52] LIU, J., IBA, H., "Selecting Informative Genes Using a Multiobjective Evolutionary Algorithm". In: *Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on*, pp. 297–302, 2002.

[53] DEB, K., REDDY, A. R., "Reliable classification of two-class cancer data using evolutionary algorithms.", *Biosystems*, v. 72, n. 1-2, pp. 111–129, Nov 2003.

[54] DEB, K., REDDY, A. R., *Classification of Two and Multi-Class Cancer Data Reliably Using Multi-Objective Evolutionary Algorithms.*, Report Report No. 2003006., Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology Kanpur, June 2003.

[55] PAUL, T., IBA, H., "Selection of the Most Useful Subset of Genes for Gene Expression-Based Classification". In: *Congress on Evolutionary Computation, 2004. CEC2004*, v. 2, pp. 2076– 2083, 2004.

[56] WAHBA, G., LIN, Y., ZHANG, H., *Advances in Large Margin Classifiers.*, chapter Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-Like Quantities., MIT Press, pp. 297–309, 2000.

[57] ALON, U., BARKAI, N., NOTTERMAN, D. A., *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.", *Proc Natl Acad Sci U S A*, v. 96, n. 12, pp. 6745–6750, June 1999.

[58] FRANK, M. B., "Northern Blot". In: Frank, M. B. (ed.), *Molecular Biology Protocols*, Oklahoma City, 1997.

[59] FRANK, M. B., "Southern Blot Protocol". In: Frank, M. B. (ed.), *Molecular Biology Protocols*, Oklahoma City, 1997.

[60] LASHKARI, D., DERISI, J., MCCUSKER, J., *et al.*, "Yeast microarrays for genome wide parallel genetic and gene expression analysis", *Proc. Natl. Acad. Sci. U.S.A.*, v. 94, pp. 13057–13062, 1997.

[61] TUSHER, V. G., TIBSHIRANI, R., CHU, G., "Significance analysis of microarrays applied to the ionizing radiation response.", *Proc Natl Acad Sci U S A*, v. 98, n. 9, pp. 5116–5121, April 2001.

[62] COBB, J. P. P., MINDRINOS, M. N. N., MILLER-GRAZIANO, C., *et al.*, "Application of genome-wide expression analysis to human health and disease", *Proc Natl Acad Sci U S A*, , March 2005.

[63] HOLLOWAY, A. J., VAN LAAR, R. K., TOTHILL, R. W., *et al.*, "Options Available–from Start to Finish–for Obtaining Data from DNA Microarrays II", *Nat Genet*, v. 32(Suppl.), pp. 481–489, 2002.

[64] HARDIMAN, G., "Microarray Platforms–Comparisons and Contrasts", *Pharmacogenomics*, v. 5, pp. 487–502, 2004.

[65] SCHENA, M., SHALON, D., DAVIS, R. W., *et al.*, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science*, v. 270, n. 5235, pp. 467–470, October 1995.

[66] DERISI, J., PENLAND, L., BROWN, P., *et al.*, "Use of a cDNA microarray to analyze gene expression patterns in human cancer", *Nat Genet*, v. 14, n. 4, pp. 457–460, 1996.

[67] LOCKHART, D. J., DONG, H., BYRNE, M. C., *et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays.", *Nat Biotechnol*, v. 14, n. 13, pp. 1675–1680, December 1996.

[68] CHURCHILL, G. A., "Fundamentals of Experimental Design for cDNA Microarrays", *Nat Genet*, v. 32(Suppl.), pp. 490–495, 2002.

[69] SCHULZE, A., DOWNWARD, J., "Navigating gene expression using microarrays–a technology review.", *Nat Cell Biol*, v. 3, n. 8, August 2001.

[70] BENNETT, H., DERISI, J., V.IYER, *et al.*, *Cold Spring Harbor Laboratory Microarray Course manual*, Report, Brown and DeRisi labs., 2001.

[71] BEAUCAGE, S., IYER, R., "Advances in the synthesis of oligonucleotides by the phosphoramidite approach", *Tetrahedron*, v. 48, pp. 2223–2311, 1992.

[72] BEAUCAGE, S., IYER, R., "The synthesis of modified oligonucleotides by the phosphoramidite approach and their applications", *Tetrahedron*, v. 49, n. 12, pp. 6123–6194, 1993.

[73] AFFYMETRIX, *GeneChip Analysis Suite: User Guide, version 3.3*. Affymetrix, 1999.

[74] AFFYMETRIX, *GeneChip Expression Analysis*. *GeneChip Expression Analysis*, 2000.

[75] KUO, R. J., WU, P., WANG, C. P., "An intelligent sales forecasting system through integration of artificial neural networks and fuzzy neural networks with fuzzy weight elimination", *Neural Networks*, v. 15, pp. 909–925, 2002.

[76] TAN, P. K., DOWNEY, T. J., SPITZNAGEL, E. L., J., *et al.*, "Evaluation of Gene Expression Measurements from Commercial Microarray Platforms", *Nucleic Acids Res*, v. 31, pp. 5676–5684, 2003.

[77] WOO, Y., AFFOURTIT, J., DAIGLE, S., *et al.*, "A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms", *J. Biomol. Tech.*, v. 15, pp. 276–284, 2004.

[78] DUFFY, M. J., "Clinical Uses of Tumor Markers: A Critical Review", *Rev. Clin. Lab. Sci.*, v. 38, pp. 225–262, 2001.

[79] NUTT, C., MANI, D., BETENSKY, R., *et al.*, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification", *Cancer Res*, v. 63, pp. 1602–1607, 2003.

[80] SU, A. I., WELSH, J. B., SAPINOSO, L. M., *et al.*, "Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures", *Cancer Res*, v. 61, pp. 7388–7393, 2001.

[81] GIORDANO, T. J., SHEDDEN, K. A., SCHWARTZ, D. R., *et al.*, "Organ-Specific Molecular Classification of Primary Lung, Colon, and Ovarian Adenocarcinomas Using Gene Expression Profiles", *Am. J. Pathol.*, v. 159, pp. 1231–1238, 2001.

[82] GRANT, G. R., MANDUCHI, E., STOECKERT, C. J., *Methods of Microarray Data Analysis*, chapter Using nonparametric methods in the context of multiple testing to identify differentially expressed genes, Boston, Kluwer Academic Publishers, pp. 37–56, 2001.

[83] CULHANE, A. C., PERRIERE, G., CONSIDINE, E. C., *et al.*, "Between-Group Analysis of Microarray Data", *Bioinformatics*, v. 18, pp. 1600–1608, 2002.

[84] NGUYEN, D., ARPAT, A., WANG, N., *et al.*, "DNA microarray experiments: biological and technological aspects", *Biometrics*, , n. 701–17, pp. 58, 2002.

[85] BLOOM, G., YANG, I. V., BOULWARE, D., *et al.*, "Multi-platform, multi-site, microarray-based human tumor classification.", *Am J Pathol*, v. 164, n. 1, pp. 9–16, Jan 2004.

[86] ROSENWALD, A., WRIGHT, G., CHAN, W. C., *et al.*, "The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma", *N Engl J Med*, v. 346, pp. 1937–1947, 2002.

[87] SHIPP, M. A., ROSS, K. N., TAMAYO, P., *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.", *Nat Med*, v. 8, n. 1, pp. 68–74, January 2002.

[88] LOSSOS, I. S., CZERWINSKI, D. K., ALIZADEH, A. A., *et al.*, "Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes", *N Engl J Med*, v. 350, pp. 1828–1837, 2004.

[89] DAVE, S. S., WRIGHT, G., TAN, B., *et al.*, "Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumorinfiltrating Immune Cells", *N Engl J Med*, v. 351, pp. 2159–2169, 2004.

[90] VAN 'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., *et al.*, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer", *Nature*, v. 415, pp. 530–536, 2004.

[91] WANG, Y., "Gene expression-driven diagnostics and pharmacogenomics in cancer.", *Curr Opin Mol Ther*, v. 7, n. 3, pp. 246–250, Jun 2005.

[92] SORLIE, T., PEROU, C. M., TIBSHIRANI, R., *et al.*, "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications", *Proc. Natl. Acad. Sci. USA*, v. 98, pp. 10869–10874, 2001.

[93] AHR, A., KARN, T., SOLBACH, C., *et al.*, "Identification of High Risk Breast-Cancer Patients by Gene Expression Profiling", *Lancet*, v. 359, pp. 131–132, 2002.

[94] RAMASWAMY, S., ROSS, K. N., LANDER, E. S., *et al.*, "A Molecular Signature of Metastasis in Primary Solid Tumors", *Nat Genet*, v. 33, pp. 49–54, 2003.

[95] GLINSKY, G. V., HIGASHIYAMA, T., GLINSKII, A. B., "Classification of Human Breast Cancer Using Gene Expression Profiling as a Component of the Survival Predictor Algorithm", *Clin. Cancer Res.*, v. 10, pp. 2272–2283, 2004.

[96] VAN DE VIJVER, M. J., HE, Y. D., VAN 'T VEER, L. J., *et al.*, "A Gene Expression Signature as a Predictor of Survival in Breast Cancer", *N Engl J Med*, v. 347, pp. 1999–2009, 2002.

[97] DHANASEKARAN, S. M., BARRETTE, T. R., GHOSH, D., *et al.*, "Delineation of Prognostic Biomarkers in Prostate Cancer", *Nature*, v. 412, pp. 822–826, 2001.

[98] SINGH, D., FEBBO, P., ROSS, K., *et al.*, "Gene expression correlates of clinical prostate cancer behavior", *Cancer Cell*, v. 1, pp. 203–209, 2002.

[99] GLINSKY, G. V., GLINSKII, A. B., STEPHENSON, A. J., *et al.*, "Gene Expression Profiling Predicts Clinical Outcome of Prostate Cancer", *J Clin Invest*, v. 113, pp. 913–923, 2004.

[100] MOCH, H., SCHRAML, P., BUBENDORF, L., *et al.*, "[Identification of prognostic parameters for renal cell carcinoma by cDNA arrays and cell chips]", *Verh Dtsch Ges Pathol*, v. 83, pp. 225–232, 1999.

[101] WANG, Y., JATKOE, T., ZHANG, Y., *et al.*, "Gene Expression Profiles and Molecular Markers to Predict Recurrence of Dukes' B Colon Cancer", *J. Clin. Oncol.*, v. 22, pp. 1564–1571, 2004.

[102] ESCHRICH, S., YANG, I., BLOOM, G., *et al.*, "Molecular staging for survival prediction of colorectal cancer patients.", *J Clin Oncol*, v. 23, n. 15, pp. 3526–3535, May 2005.

[103] IIZUKA, N., OKA, M., YAMADA-OKABE, H., *et al.*, "Oligonucleotide microarray for preditction of early intra-hepatic recurrence of hepatocellular carcinoma after curative resection", *The Lancet*, v. 361, pp. 923–929, 2003.

[104] ROEPMAN, P., WESSELS, L. F., KETTELARIJ, N., *et al.*, "An expression profile for diagnosis of lymph node metástases from primary head and neck squamous cell carcinomas.", *Nat Genet*, v. 37, n. 2, pp. 182–186, February 2005.

[105] BEER, D. G., KARDIA, S. L., HUANG, C. C., *et al.*, "Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma", *Nat. Med.*, v. 8, pp. 816–824, 2002.

[106] KIHARA, C., TSUNODA, T., TANAKA, T., *et al.*, "Prediction of Sensitivity of Esophageal Tumors to Adjuvant Chemotherapy by cDNA Microarray Analysis of Gene-Expression Profiles", *Cancer Res*, v. 61, pp. 6474–6479, 2001.

[107] FERNANDEZ-TEIJEIRO, A., BETENSKY, R. A., STURLA, L. M., *et al.*, "Combining Gene Expression Profiles and Clinical Parameters for Risk Stratification in Medulloblastomas", *J. Clin. Oncol.*, v. 22, pp. 994–998, 2004.

[108] ARMITAGE, J. O., WEISENBURGER, D., "New Approach to Classifying Non-Hodgkin's Lynphomas: Clinical Features of the Major Histologic Types. Non-Hodgking's Lynphoma Classification Project", *J. Clin. Oncol.*, v. 8, pp. 2780–2795, 1998.

[109] HORNING, S. J., "Follicular Lymphoma: Have We Made Any Progress?", *Ann. Oncol.*, v. 11, n. Suppl. 1, pp. 23–27, 2000.

[110] PEROU, C. M., JEFFREY, S. S., VAN DE RIJN, M., *et al.*, "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers", *Proc Natl Acad Sci U S A*, v. 96, n. 16, pp. 9212–9217, Aug 3 1999.

[111] EIN-DOR, L., KELA, I., GETZ, G., *et al.*, "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?", *Bioinformatics*, v. 21, n. 2, pp. 171–178, 2005.

[112] SCHERF, U., ROSS, D. T., WALTHAM, M., *et al.*, "A gene expression database for the molecular pharmacology of cancer.", *Nat Genet*, v. 24, pp. 236–244, march 2000.

[113] DAN, S., TSUNODA, T., KITAHARA, O., *et al.*, "An Integrated Database of Chemosensitivity to 55 Anticancer Drugs and Gene Expression Profiles of 39 Human Cancer Cell Lines", *Cancer Res*, v. 62, pp. 1139–1147, 2002.

[114] CHANG, J. C., WOOTEN, E. C., TSIMELZON, A., *et al.*, "Gene Expression Profiling for the Prediction of Therapeutic Response to Docetaxel in Patients with Breast Cancer", *Lancet*, v. 362, pp. 362–369, 2003.

[115] AYERS, M., SYMMANS, W. F., STEC, J., *et al.*, "Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer", *J. Clin. Oncol.*, v. 22, pp. 2284–2293, 2004.

[116] JANSEN, M. P., FOEKENS, J. A., VAN STAVEREN, I. L., *et al.*, "Molecular Classification of Tamoxifen-Resistant Breast Carcinomas by Gene Expression Profiling", *J. Clin. Oncol.*, v. 23, pp. 732–740, 2005.

[117] HOLLEMAN, A., CHEOK, M. H., DEN BOER, M. L., *et al.*, "Gene Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment", *N Engl J Med*, v. 351, pp. 533–542, 2004.

[118] HOFMANN, W. K., DE VOS, S., ELASHOFF, D., *et al.*, "Relation Between Resistance of Philadelphia-Chromosome-Positive Acute Lymphoblastic Leukaemia to the Tyrosine Kinase Inhibitor STI571 and Gene-Expression Profiles: A Gene-Expression Study", *Lancet*, v. 359, pp. 481–486, 2002.

[119] MAYR, E., *Toward a New Philosophy of Biology: Observations of an Evolutionist*. Cambridge, MA, USA, Belknap, 1987.

[120] HOLLAND, J. H., "Outline for a logical theory of adaptive systems", *J. ACM*, v. 9, pp. 297–314, 1962.

[121] BAKER, J., "Reducing Bias and Inefficiency in the Selection Algorithms". In: *ICGA1*, pp. 101–111, 1985.

[122] BLICKLE, T., THIELE, L., *A Comparison of Selection Schemes used in Genetic Algorithms*, Technical Report 11, Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Thechnology (ETH), Zürich, Switzerland, 1995. 2nd Edition.

[123] WHITLEY, D., "The GENITOR Algorithm and Selection Pressure: Why Rank-Based Allocation of Reproducive Trials is Best". In: *ICGA3*, pp. 116–121, 1989.

[124] BÄCK, T., HOFFMEISTER, F., "Extended Selection Mechanisms in Genetic Algorithms". In: *ICGA4*, pp. 92–99, 1991.

[125] GOLDBERG, D., DEB, K., "A Comparative Analysis of Selection Schemes Used in Genetic Algorithms". In: *FGA1*, pp. 69–93, 1991.

[126] BOOKER, L., *Improving Search in Genetic Algorithms*, Morgan Kaufmann Publishers, pp. 61–73, 1987.

[127] SPEARS, W. M., DE JONG, K. A., "On the Virtues of Parameterised Uniform Crossover". In: Forrest, S. (ed.), *Proceedings of the Fourth International Conference on Genetic Algorithms*, pp. 230–236, San Mateo, California, USA, 1991.

[128] SYSWERDA, G., "Uniform crossover in genetic algorithms". In: *ICGA3*, pp. 2–9, 1989.

[129] SPEARS, W., DE JONG, K. A., *An Analysis of Multi-Point Crossover*, Morgan Kaufmann Publishers, pp. 301–315, 1991.

[130] CARUANA, R. A., ESHELMANN, L., SCHAFFER, J., "Representation and Hidden Bias II: Eliminating Defining Length Bias in Genetic Search Via Shuffle Crossover". In: Sridharan, N. (ed.), *Eleventh International Join Conference on Artificial Intelligence*, pp. 750–755, San Mateo, California, USA, 1989.

[131] MATHIAS, K., WHITLEY, L. D., "Changing representations during search: a comparative study of delta coding". In: Bäck, T., Fogel, D. B., Michalewicz, Z. (eds.), *Evolutionary Computation 2: Advanced Algorithms and Operators*, 1994.

[132] FONSECA, C. M., FLEMING, P. J., "An Overview of Evolutionary Algorithms in Multiobjective Optimization", *Evolutionary Computation*, v. 3, n. 1, pp. 1–16, 1995.

[133] ZITZLER, E., DEB, K., THIELE, L., "Comparison of multiobjective evolutionary algorithms: empirical results.", *Evolutionary Computation*, v. 8, n. 2, pp. 173–195, 2000.

[134] SRINIVAS, N., DEB, K., "Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms", *Evolutionary Computation*, v. 2, n. 3, pp. 221–248, 1994.

[135] FONSECA, C. M., *Multiobjective Genetic Algorithms with Application to Control Engineering Problems.* Ph.d. thesis, Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, U.K., 1995.

[136] MAHFOUD, S. W., *Niching methods for genetic algorithms*. Ph.D. dissertation, Urbana, IL, USA, 1995.

[137] GOLDBERG, D. E., RICHARDSON, J., "Genetic algorithms with sharing for multimodal function optimization." In: *Proceedings of the Second International Conference on Genetic Algorithms*, pp. 41–49, Cambridge, MA., 1987.

[138] DUDOIT, S., YANG, Y., CALLOW, M., *et al.*, "Statistical methods for identifying expressed genes in replicated cDNA experiments", *Stat. Sinica*, , n. 12, pp. 111–140, 2002.

[139] PAN, W., "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.", *Bioinformatics*, v. 18, n. 4, pp. 546–554, April 2002.

[140] PARMIGIANI, G., *The Analysis of Gene Expression Data*. Springer, April 2003.

[141] DRAGHICI, S., KUKLIN, A., HOFF, B., *et al.*, "Experimental Design, Analysis of Variance and Slide Quality Assessment in Gene Expression Arrays", *Current Opinion in Drug Discovery & Development*, v. 4, n. 3, pp. 332–337, 2001.

[142] KERR, M., LEITER, E., CHURCHILL, G., *Analysis of a designed microaray experiment*, Technical report, The Jackson Laboratory, 2000.

[143] MIKE, W., NEVINS, J., MARKS, J., *et al.*, "Bayesian Regression Analysis in the 'Large p, Small n' Paradigm with Application in DNA MIcroarray Studies", 2000.

[144] FRIEDMAN, N., LINIAL, M., NACHMAN, I., *et al.*, "Using Bayesian networks to analyze expression data.", *J Comput Biol*, v. 7, n. 3-4, pp. 601–620, 2000.

[145] DROR, R. O., "Noise Models in Gene Array Analysis".

[146] LONG, A. D., MANGALAM, H. J., CHAN, B. Y., *et al.*, "Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework".

[147] SMYTH, G. K. K., MICHAUD, J., SCOTT, H. S. S., "Use of within-array replicate spots for assessing differential expression in microarray experiments.", *Bioinformatics*, v. 21, n. 9, pp. 2067–75, January 2005.

[148] FURLANELLO, C., SERAFINI, M., MERLER, S., *et al.*, "An Accelerated Procedure for Recursive Feature Ranking on Microarray Data", *Neural Networks*, v. 16, pp. 641–648, 2003.

[149] CHO, J.-H., LEE, D., PARK, J. H., *et al.*, "Gene Selection and Classification from Microarray Data Using Kernel Machine", *FEBS Lett*, v. 571, pp. 93–98, 2004.

[150] WAHDE, M., SZALLASI, Z., "Improving the prediction of the clinical outcome of breast cancer using evolutionary algorithms", *Soft Comput.*, v. 10, n. 4, pp. 338–345, 2006.

[151] DEUTSCH, J. M., "Evolutionary algorithms for finding optimal gene sets in microarray prediction", *Bioinformatics*, v. 19, n. 1, pp. 45–52, 2003.

[152] ROSS, D., SCHERF, U., EISEN, M., *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines", *Nat Genet.*, v. 24, n. 3, pp. 227–235., Mar 2000.

[153] CHAPELLE, O., VAPNIK, V., BOUSQUET, O., *et al.*, "Choosing Multiple Parameters for Support Vector Machines.", *Machine Learning*, v. 46, n. 1, pp. 131–159, 2002.

[154] WANG, Y., MAKEDON, F., "Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification Using Microarray Data". In: *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pp. 497–498, Washington, DC, USA, 2004.

[155] DEB, K., AGRAWAL, S., PRATAB, A., *et al.*, "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II". In: Schoenauer, M., Deb, K., Rudolph, G., *et al.* (eds.), *Proceedings of the Parallel Problem Solving from Nature VI Conference*, pp. 849–858, Paris, France, 2000.

[156] CORTES, C., VAPNIK, V., "Support Vector Networks", *Machine Learning*, v. 20, pp. 273–297, 1995.

[157] BUTTE, A., "The use and analysis of microarray data.", *Nat Rev Drug Discov*, v. 1, n. 12, pp. 951–960, December 2002.

[158] BROWN, M., GRUNDY, W., LIN, D., *et al.*, *Support vector machine classification of microarray gene expression data*, Technical Report UCSC-CRL 99-09, Department of Computer Science, University California Santa Cruz, Santa Cruz, CA, 1999.

[159] BROWN, M., GRUNDY, W., LIN, D., *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines", *Proc Natl Acad Sci U S A.*, v. 97, n. 1, pp. 262–267, Jan 4 2000.

[160] EISEN, M., SPELLMAN, P., BROWN, P., *et al.*, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci. U.S.A.*, v. 95, pp. 14863–14868, 1998.

[161] TIBSHIRANI, R., HASTIE, T., EISEN, M., *et al.*, *Clustering methods for the analysis of dna microarray data*, Report, Department of Health Reseach and Policy, Department of Statistics, Department of Genetics and Department of Biochemistry. Stanford University, October, 15 1999.

[162] CHENG, Y., CHURCH, G. M., "Biclustering of Expression Data". In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103, 2000.

[163] POCHET, N., SMET, F. D., SUYKENS, J. A. K., *et al.*, "Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.", *Bioinformatics*, v. 20, n. 17, pp. 3185–3195, Nov 2004.

[164] YEANG, C. H., RAMASWAMY, S., TAMAYO, P., *et al.*, "Molecular classification of multiple tumor types.", *Bioinformatics*, v. 17 Suppl 1, pp. S316–S322, 2001.

[165] MARKOWETZ, F., HEYDEBRECK, A. V., "Class discovery in gene expression data: characterizing splits by support vector machines". In: *Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation 2002*, pp. 662–669, Gesellschaft für Klassifikation (GfKl), 2002.

[166] LEE, Y., LEE, C.-K., "Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data", *Bioinformatics*, v. 19, n. 9, pp. 1132–1139, 2003.

[167] POHLHEIM, H., *GEATbx: Genetic and Evolutionary Algorithm Toolbox for Use with Matlab*. Www.Geatbx.Com, 1994–2007.

[168] CHANG, C.-C., LIN, C.-J., *LIBSVM: a library for support vector machines. LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[169] TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., *et al.*, "Missing value estimation methods for DNA microarrays.", *Bioinformatics*, v. 17, n. 6, pp. 520–525, June 2001.

[170] BALUJA, S., *Population Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitve Learning*, Technical Report CMU-CS-94-163, Carnegie Mellon University, Pittsburg, Pensylvania, 1994.

[171] YAO, X., "Evolutionary artificial neural networks". In: Kent, A., Williams, J. G. (eds.), *Encyclopedia of Computer Science and Technology*, v. 33, Marcel Dekker Inc., pp. 137–170, 1995.

[172] YAO, X., "Evolving Artificial Neural Networks", *PIEEE: Proceedings of the IEEE*, v. 87, pp. 1423–1447, 2003.

[173] RIFKIN, R., MUKHERJEE, S., TAMAYO, P., *et al.*, "An Analytical Method for Multiclass Molecular Cancer Classification", *SIAM REVIEW*, v. 45, n. 4, pp. 706–723, 2003.

[174] BRAGA-NETO, U., DOUGHERTY, E., "Is cross-validation valid for small-sample microarray classification?", *Bioinformatics*, v. 20, n. 3, pp. 374–380, August 2003.

[175] BRAGA-NETO, U., DOUGHERTY, E., "Bolstered error estimation", *Pattern Recognition*, v. 37, pp. 1627–1281, 2004.

[176] BATTITI, R., "Using Mutual Information for Selecting Features in supervised Neural Net Learning", *IEEE Trans. on Neural Networks*, v. 5, n. 4, pp. 537–550, 1994.

[177] KWAK, N., CHOI, C.-H., "Input Feature Selection by Mutual Information Based on Parzen Window", *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 24, n. 12, pp. 1667–1671, 2002.

[178] DING, C., PENG, H., "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". In: *CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics*, p. 523, Washington, DC, USA, 2003.

[179] LONG, F., DING, C., "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 27, n. 8, pp. 1226–1238, 2005. Member-Hanchuan Peng.

[180] TORKKOLA, K., “Feature Extraction by Non-Parametric Mutual Information Maximization”, *Journal of Machine Learning Research*, v. 3, pp. 1415–1438, Mar 2003.

Apêndice A

Resultados Adicionais

A.1 Leucemia Aguda

Tabela A.1: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 1.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	85,29	10	32
	100	100	0	85,29	10	32
	100	100	0	85,29	10	32
	100	100	0	85,29	10	32
	100	100	0	85,29	10	32
	100	100	0	85,29	10	32
	100	97,36	0	85,29	10	30
	100	97,36	0	85,29	10	30
	100	100	0	85,29	10	32
	100	94,73	0	76,47	10	29
	100	100	0	79,41	10	32
	100	94,73	0	82,35	10	29
média $\pm$ std	100 ± 0	$98,68 \pm 2,09$	0 ± 0	$83,82 \pm 2,94$	10 ± 0	$31,16 \pm 1,26$

Tabela A.2: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 2.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	85,29	0,1	35
	100	100	0	85,29	0,1	35
	100	94,73	0	76,47	100	30
	100	94,73	0	76,47	100	30
	100	100	0	85,29	0,1	35
	100	100	0	76,47	1	35
	100	97,36	0	82,35	10	31
	100	100	0	73,52	100	35
	100	100	0	76,47	0,1	35
média $\pm$ std	100 ± 0	$98,53 \pm 2,32$	0 ± 0	$79,73 \pm 4,75$	$34,6 \pm 49,15$	$33,44 \pm 2,35$

Tabela A.3: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 3.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	97,36	0	61,76	100	30
	100	97,36	0	61,76	100	30
	100	100	0	85,29	100	39
média $\pm$ std	100 ± 0	$98,24 \pm 1,51$	0 ± 0	$69,6 \pm 13,58$	100 ± 0	$33 \pm 5,19$

Tabela A.4: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 4.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	94,73	0	79,41	10	28
	100	94,73	0	79,41	10	28
	100	100	0	73,52	1	30
média $\pm$ std	100 ± 0	$96,49 \pm 3,03$	0 ± 0	$77,45 \pm 3,39$	$7 \pm 5,19$	$28,66 \pm 1,15$

Tabela A.5: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 5.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	94,73	0	70,58	0,1	30
	100	94,73	0	70,58	0,1	30
	100	97,36	0	88,23	1	31
	100	100	0	79,41	1	33
	100	100	0	82,35	1	33
média $\pm$ std	100 ± 0	$97,36 \pm 2,63$	0 ± 0	$78,23 \pm 7,66$	$0,64 \pm 0,49$	$31,4 \pm 1,51$

Tabela A.6: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 6.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^o de Variáveis**
	100	100	0	88,23	10	29
	100	100	0	88,23	10	29
	100	97,36	0	85,29	10	28
média $\pm$ std	100 ± 0	$99,12 \pm 1,51$	0 ± 0	$87,25 \pm 1,69$	10 ± 0	$28,66 \pm 0,57$

Tabela A.7: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 7.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	94,73	0	73,52	10	30
	100	94,73	0	73,52	10	30
	100	100	0	76,47	0,1	32
média $\pm$ std	100 ± 0	$96,49 \pm 3,03$	0 ± 0	$74,5 \pm 1,69$	$6,7 \pm 5,71$	$30,66 \pm 1,15$

Tabela A.8: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 8.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	82,35	0,1	30
	100	100	0	82,35	0,1	30
média $\pm$ std	100 ± 0	100 ± 0	0 ± 0	$82,35 \pm 0$	$0,1 \pm 0$	30 ± 0

Tabela A.9: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 9.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	97,36	0	76,47	100	30
	100	97,36	0	76,47	100	30
	100	100	0	91,17	100	34
média $\pm$ **std**	100 ± 0	$98,68 \pm 2,09$	0 ± 0	$83,82 \pm 2,94$	10 ± 0	$31,16 \pm 1,26$

Tabela A.10: Resultados de treino com o conjunto de dados de Leucemia Aguda em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes AML e ALL, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 10.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	70,58	10	30
	100	100	0	70,58	10	30
	100	94,73	0	73,52	10	26
	100	100	0	82,35	10	30
	100	100	0	76,47	10	30
média $\pm$ std	100 ± 0	$98,94 \pm 2,35$	0 ± 0	$74,7 \pm 4,92$	10 ± 0	$29,2 \pm 1,78$

Tabela A.11: Resultado do emprego exclusivo da eliminação recursiva de atributos (RFE–SVM) para o conjunto de dados de Leucemia Aguda segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} de Atributos	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	94,73	94,11	94,44
2	100	94,11	97,22
3	100	97,05	98,61
4	100	97,05	98,61
5	100	85,29	93,05
6	100	94,11	97,22
7	100	94,11	97,22
8	100	94,11	97,22
9	100	94,11	97,22
10	100	94,11	97,22
11	100	94,11	97,22
12	100	97,05	98,61
13	100	97,05	98,61
14	100	97,05	98,61
15	100	97,05	98,61
16	100	97,05	98,61
17	100	94,11	97,22
18	100	94,11	97,22
19	100	94,11	97,22
20	100	94,11	97,22
21	100	91,17	95,83
22	100	91,17	95,83
23	100	91,17	95,83
24	100	91,17	95,83
25	100	91,17	95,83

Tabela A.12: Assinatura gênica do melhor classificador encontrado para o conjunto de dados de Leucemia Aguda pelo uso exclusivo da RFE–SVM.

Índice	**Descrição**	**Símbolo**
1882	CST3 Cystatin C	M27891
	(amyloid angiopathy and cerebral hemorrhage)	
1630	Inducible protein mRNA	L47738
3847	GB DEF = Homeodomain protein HoxA9 mRNA	U82759

A.2 Linfoma Difuso de Grandes Células B

Tabela A.13: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 1.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	90,9	10	33
	100	100	0	90,9	10	33
	100	100	0	90,9	10	33
	100	100	0	90,9	10	33
média $\pm$ std	100 ± 0	100 ± 0	0 ± 0	$90,9 \pm 0$	10 ± 0	33 ± 0

Tabela A.14: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 2.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	87,87	1	32
	100	100	0	87,87	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	96,82	0	90,9	100	27
	100	96,82	0	90,9	100	27
	100	100	0	90,9	1	32
	100	100	0	87,87	1	32
média $\pm$ std	100 ± 0	$99,2 \pm 1,46$	0 ± 0	$89,77 \pm 1,56$	$25,75 \pm 45,82$	$30,75 \pm 2,31$

Tabela A.15: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 3.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	95,23	0	93,93	1	29
	100	95,23	0	93,93	1	29
	100	100	0	78,78	1	37
	100	98,41	0	93,93	1	34
	100	100	0	93,93	1	37
	100	96,82	0	87,87	1	31
média $\pm$ std	100 ± 0	$97,61 \pm 2,18$	0 ± 0	$90,4 \pm 6,18$	1 ± 0	$32,83 \pm 3,71$

Tabela A.16: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 4.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	98,41	95,23	0,03	87,87	0,1	28
	100	100	0	90,90	1	35
	100	100	0	93,93	0,1	35
	100	98,41	0	90,9	0,1	29
média $\pm$ std	$98,71 \pm 0,63$	$96,03 \pm 1,73$	$0,02 \pm 0,01$	$88,63 \pm 1,74$	$0,15 \pm 0,22$	$28,93 \pm 2,37$

Tabela A.17: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 5.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	95,23	0	90,9	1	29
	100	95,23	0	90,9	1	29
	100	98,41	0	87,87	1	37
	100	100	0	93,93	1	41
média $\pm$ std	100 ± 0	$97,22 \pm 2,38$	0 ± 0	$90,9 \pm 2,47$	1 ± 0	34 ± 6

Tabela A.18: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 6.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	100	0	90,9	1	32
	100	96,82	0	87,87	1	30
	100	96,82	0	87,87	1	30
	100	100	0	90,9	1	32
	100	96,82	0	90,9	1	30
	100	100	0	90,9	1	32
	100	100	0	93,93	1	32
	100	98,41	0	90,9	1	31
	100	95,23	0	90,9	1	28
	100	98,41	0	90,9	1	31
média $\pm$ std	100 ± 0	$99,08 \pm 1,52$	0 ± 0	$90,74 \pm 1,22$	1 ± 0	$31,36 \pm 1,11$

Tabela A.19: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 7.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	98,41	96,82	0,03	90,9	0,1	31
	100	100	0	81,81	1	34
	100	95,23	0	90,9	1	30
	100	98,41	0	90,9	0,1	33
	100	98,41	0	81,81	0,1	33
	100	98,41	0	87,87	0,1	33
	100	96,82	0	90,90	0,1	32
média $\pm$ std	$99,77 \pm 0,59$	$97,73 \pm 1,54$	$0 \pm 0,01$	$87,87 \pm 4,28$	$0,35 \pm 0,43$	$32,28 \pm 1,38$

Tabela A.20: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 8.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	98,41	96,82	0,03	87,87	0,1	27
	98,41	96,82	0,03	87,87	0,1	27
	98,41	96,82	0,03	87,87	0,1	27
	100	95,23	0	93,93	0,1	28
	100	95,23	0	93,93	0,1	28
	98,41	96,82	0,03	87,87	0,1	27
	100	96,82	0	87,87	1	30
	100	98,41	0	90,9	1	31
	100	100	0	84,84	1	32
média $\pm$ std	$99,29 \pm 0,83$	$97 \pm 1,47$	$0,01 \pm 0,01$	$89,22 \pm 3,07$	$0,4 \pm 0,45$	$28,55 \pm 1,94$

Tabela A.21: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 9.

Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33
100	100	0	90,9	100	33

Tabela A.22: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 9 (cont.).

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	100	0	90,9	100	33
	100	98,41	0	90,9	100	29
média $\pm$ std	100 ± 0	$99,94 \pm 0,28$	0 ± 0	$90,9 \pm 0$	100 ± 0	$32,86 \pm 0,73$

Tabela A.23: Resultados de treino com o conjunto de dados de DLBCL em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes DLBCL e outros, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 10.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	98,41	0	90,9	10	30
	100	98,41	0	90,9	10	30
	100	100	0	90,9	10	34
média $\pm$ **std**	100 ± 0	$98,94 \pm 0,91$	0 ± 0	$90,9 \pm 0$	10 ± 0	$31,33 \pm 2,3$

Tabela A.24: Resultado do emprego exclusivo da eliminação recursiva de atributos (RFE–SVM) para o conjunto de dados de DLBCL segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} **de Atributos**	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	68,25	87,87	75
2	63,49	75,75	67,70
3	80,95	81,81	81,25
4	79,36	81,81	80,20
5	80,95	84,84	82,29
6	85,71	84,84	85,41
7	88,88	84,84	87,5
8	90,47	84,84	88,54
9	90,47	84,84	88,54
10	96,82	81,81	91,66
11	95,23	81,81	90,62
12	95,23	81,81	90,62
13	98,41	78,78	91,66
14	100	81,81	93,75
15	100	84,84	94,79
16	100	84,84	94,79
17	100	81,81	93,75
18	100	81,81	93,75
19	100	78,78	92,70
20	100	78,78	92,70
21	100	75,75	91,66
22	100	75,75	91,66
23	100	78,78	92,70
24	100	81,81	93,75
25	100	87,87	95,83

Tabela A.25: Assinatura gênica do melhor classificador encontrado para o conjunto de dados de DLBCL pelo uso exclusivo da RFE–SVM.

Índice	**Descrição**	**Símbolo**
855	Unknown; Clone=1671581	GENE2396X
449	Unknown; Clone=701622	GENE3475X
236	AIM2; Clone=683659	GENE3795X
3372	LPP; Clone=1357334	GENE3857X
2370	Similar to IL-4 receptor alpha chain; Clone=1670890	GENE1159X
1014	Unknown; Clone=1358131	GENE2102X
3235	CASPASE-5; Clone=341763	GENE144X
3141	Similar to beta2-syntrophin (SNT B2); Clone=1355039	GENE1062X
1226	Unknown UG Hs.193367 ESTs; Clone=1340261	GENE3272X
18	Similar to dead box, Y isoform (DBY); Clone=1350869	GENE3067X
3396	Unknown UG Hs.190472 ESTs; Clone=1354295	GENE1450X
3083	Unknown UG Hs.29879 ESTs; Clone=705151	GENE420X
397	Unknown UG Hs.4766 *Homo sapiens* mRNA; Clone=1307052	GENE3636X
3547	JNK1; Clone=119133	GENE3977X
487	Unknown; Clone=1289212	GENE2562X
3924	HER2; Clone=1288151	GENE1773X
108	Similar to tetraspan NET-4-1; Clone=815541	GENE3868X
2145	Unknown UG Hs.96731 huntingtin interacting protein-1-related; Clone=1334485	GENE1373X
1373	CD21, B-lymphocyte CR2-receptor; Clone=814917	GENE3393X
512	Immunoglobulin D heavy chain constant region; Clone=1370359	GENE2486X
2526	KOC; Clone=429494	GENE193X
2810	pLK; Clone=1352275	GENE703X
293	Oct-2; Clone=1319407	GENE3743X
3428	BENE; Clone=814806	GENE1435X
1452	PCM-1; Clone=345308	GENE1849X

A.3 Câncer de Cólon

Tabela A.26: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 1.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	97,5	0	68,18	100	28
	100	97,5	0	68,18	100	28
	100	100	0	68,18	1	30
média $\pm$ std	100 ± 0	$98,33 \pm 1,44$	0 ± 0	$68,18 \pm 0$	$67 \pm 57,15$	$28,66 \pm 1,15$

Tabela A.27: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 2.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	100	0	72,72	100	28
	100	100	0	72,72	100	28
média $\pm$ std	100 ± 0	100 ± 0	0 ± 0	$72,72 \pm 0$	100 ± 0	28 ± 0

Tabela A.28: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 3.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	72,72	100	29
	100	100	0	72,72	100	29
	100	97,5	0	68,18	1	27
	100	100	0	68,18	1	29
média $\pm$ std	100 ± 0	$99,37 \pm 1,25$	0 ± 0	$70,45 \pm 2,62$	$50,5 \pm 57,15$	$28,5 \pm 1$

Tabela A.29: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 4.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	77,27	10	33
	100	100	0	77,27	10	33
	100	97,5	0	81,81	1	30
	100	97,5	0	81,81	1	30
	100	100	0	77,27	10	33
	100	100	0	77,27	10	33
média $\pm$ std	100 ± 0	$99,16 \pm 1,29$	0 ± 0	$78,78 \pm 2,34$	$7 \pm 4,64$	$32 \pm 1,54$

Tabela A.30: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 5.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	95	0	68,18	100	27
	100	95	0	68,18	100	27
	100	97,5	0	68,18	0,1	30
	100	95	0	68,18	0,1	27
	100	100	0	77,27	0,1	32
média $\pm$ std	100 ± 0	$96,5 \pm 2,23$	0 ± 0	$70 \pm 4,06$	$40,06 \pm 54,71$	$28,6 \pm 2,3$

Tabela A.31: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 6.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	100	0	68,18	10	25
	100	100	0	68,18	10	25
média $\pm$ std	100 ± 0	100 ± 0	0 ± 0	$68,18 \pm 0$	10 ± 0	25 ± 0

Tabela A.32: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 7.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	N^{o} **de Variáveis**
	100	100	0	72,72	1	30
	100	100	0	72,72	1	30
média $\pm$ std	100 ± 0	100 ± 0	0 ± 0	$72,72 \pm 0$	1 ± 0	30 ± 0

Tabela A.33: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 8.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	97,5	0	77,27	1	30
	100	97,5	0	77,27	1	30
	100	100	0	72,72	1	31
	100	97,5	0	77,27	1	30
	100	95	0	63,63	1	28
	100	97,5	0	72,72	1	30
média $\pm$ std	100 ± 0	$97,5 \pm 1,58$	0 ± 0	$73,48 \pm 5,31$	1 ± 0	$29,83 \pm 0,98$

Tabela A.34: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 9.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	97,5	0	77,27	0,1	25
	100	97,5	0	77,27	0,1	25
	100	100	0	68,18	0,1	29
média $\pm$ std	100 ± 0	$98,33 \pm 1,44$	0 ± 0	$74,24 \pm 5,24$	$0,1 \pm 0$	$26,33 \pm 2,3$

Tabela A.35: Resultados de treino com o conjunto de dados de Câncer de Cólon em função dos valores das acurácias de treino, validação cruzada *leave-one-out* (LOOCV), diferença entre a razão de erro do classificador entre as classes normal e câncer, acurácia de teste, parâmetro de regularização C e número de variáveis selecionadas na simulação 10.

	Acurácia de Treino	**Acurácia LOOCV**	**Dif. Razão de Erro**	**Acurácia de Teste**	**C**	**N^{o} de Variáveis**
	100	97,5	0	77,27	0,1	29
	100	97,5	0	77,27	0,1	29
	97,5	95	0,03	77,27	0,1	28
	100	100	0	68,18	0,1	31
média $\pm$ std	$99,37 \pm 1,25$	$97,5 \pm 2,04$	$0 \pm 0,01$	$75 \pm 4,54$	$0,1 \pm 0$	$29,25 \pm 1,25$

Tabela A.36: Resultado do emprego exclusivo da eliminação recursiva de atributos (RFE–SVM) para o conjunto de dados de Câncer de Cólon segundo o número de atributos e as acurácias de treino, teste e total.

N^{o} **de Atributos**	**Acurácia de Treino** (%)	**Acurácia de Teste** (%)	**Acurácia Total** (%)
1	65	63,63	64,51
2	65	63,63	64,51
3	65	63,63	64,51
4	65	63,63	64,51
5	80	59,09	72,58
6	77,5	54,54	69,35
7	77,5	54,54	69,35
8	77,5	59,09	70,96
9	77,5	59,09	70,96
10	77,5	59,09	70,96
11	80	59,09	72,58
12	97,5	77,27	90,32
13	92,5	77,27	87,09
14	100	86,36	95,16
15	100	77,27	91,93
16	100	86,36	95,16
17	97,5	77,27	90,32
18	97,5	77,27	90,32
19	97,5	72,72	88,70
20	100	86,36	95,16
21	97,5	81,81	91,93
22	97,5	81,81	91,93
23	97,5	77,27	90,32
24	100	77,27	91,93
25	97,5	81,81	91,93

Tabela A.37: Assinatura gênica do melhor classificador encontrado para o conjunto de dados de Câncer de Cólon pelo uso exclusivo da RFE–SVM.

Índice	**Descrição**	**Símbolo**
536	lymphocyte antigen (HLA-G2.1) mRNA; *Homo sapiens*	M90684
827	Tubulin Alpha-5 Chain; *Gallus gallus*	T90759
744	PCR amplified genomic DNA for GPE (exon 1); *H. sapiens*	X53004
471	Phosphoglycerate Mutase, Brain; *H. sapiens*	J04173
433	Glutathione Peroxidase; *H. sapiens*	T52343
162	SOX-4 Protein; *H. sapiens*	T90774
1369	Probable Serine/Threonine-Protein Kinase C16C9.07; *Schizosaccharomyces pombe*	H53092
24	Thymosin Beta-4; *H. sapiens*	T59954
528	RD Protein; *H. sapiens*	R20804
1565	Genome Polyprotein; *Langat virus*	H08351
1855	mRNA KIAA0077; *H. sapiens*	D38521
1062	mRNA for protein p68Human mRNA for protein p68; *H. sapiens*	Y00097
602	tetracycline transporter-like protein mRNA; *H. sapiens*	H28711
44	Ubiquitin; *H. sapiens*	T88723