CLASSIFICAÇÃO DE ONCOGENES MEDIDOS POR *MICROARRAY* UTILIZANDO ALGORITMOS GENÉTICOS

Laurence Rodrigues do Amaral*, Geraldo Sadoyama†, Foued Salmen Espindola‡, Gina Maira B. Oliveira*

*Av. João Naves de Ávila, 2160
Universidade Federal de Uberlândia
Laboratório de Inteligência Artificial
Uberlândia, MG, Brasil

†Av. Arthur Botelho, S/N
Centro Universitário do Cerrado-Patrocínio
Laboratório de Imunologia, Genética e Microbiologia
Patrocínio, MG, Brasil

‡Av. Pará, 1720 Bloco 2E39A
Universidade Federal de Uberlândia
Instituto de Genética e Bioquímica
Laboratório de Bioquímica e Biologia Molecular
Uberlândia, MG, Brasil

Emails: lramaral@pos.facom.ufu.br,laurence@unicerp.edu.br, geraldosadoyama@unicerp.edu.br, foued@ufu.br, gina@facom.ufu.br

Resumo— Técnicas de Inteligência Artificial (IA) têm se tornado cada vez mais importantes na solução de problemas biológicos. Neste artigo, utilizamos um Algoritmo Genético (AG) na busca de regras de alto nível do tipo IF-THEN. Este AG foi aplicado na mineração de regras de classificação em uma base de dados de expressão gênica de células cancerígenas, advindas de experimentos de *microarray*. O objetivo dessa mineração é descobrir relações entre os níveis de expressões gênicas e os nove tipos de classes de câncer analisados.

Palavras-chave— Bioinformática, expressão gênica, algoritmos genéticos, oncogenes, *data mining*

1 Introdução

Uma das áreas em que a aplicação de técnicas computacionais inteligentes tem se mostrado mais promissora é a Biologia Molecular (Setúbal and Meidanis, 1997).

Devido à grande quantidade e complexidade da informação, as ferramentas baseadas na computação convencional têm se mostrado limitadas na abordagem de problemas biológicos complexos. Uma das explicações para essa dificuldade é a ineficiência das ferramentas convencionais em lidar com grandes volumes de dados. Técnicas advindas da Inteligência Artificial (IA), tais como, os algoritmos genéticos e as redes neurais artificiais, são cada vez mais empregadas para tratar problemas em Biologia Molecular. A aplicabilidade dessas técnicas advém de sua capacidade de aprender automaticamente a partir de grandes volumes de dados e produzir hipóteses úteis (Baldi and Brunak, 2001).

Um fragmento de DNA pode conter diversos genes. A propriedade mais importante dos genes está no fato de que eles contêm o código genético para a expressão do mRNA (RNA mensageiro) que será traduzido em proteínas, componentes estes, essenciais a todo ser vivo (Souto et al., 2003). As proteínas são polipeptídeos compostas por conjuntos de aminoácidos. Estes aminoácidos são representados por trincas (códons) de nucleotídeos (Adenina - A, Uracila - U, Citosina - C e Guanina - G) no DNA. O processo pelo qual as seqüências de nucleotídeos dos genes são interpretados na produção de proteínas é denominado expressão gênica (Souto et al., 2003). Mensurar e analisar informações de expressão gênica é de grande interesse para as Ciências Biológicas. Esse tipo de análise pode fornecer informações importantes sobre as funções de uma célula, uma vez que as mudanças na fisiologia de um organismo são geralmente acompanhadas por mudanças nos padrões de expressão dos genes (Alberts et al., 1997). Uma das técnicas mais difundidas para esta medição são os *Microarrays* de DNA (Velculescu et al., 1995; Freeman et al., 1999).

Diferentes técnicas de IA foram aplicadas na análise de dados de expressão gênica, tais como: redes neurais artificiais (Xu et al., 2002; Khan et al., 2001), *support vector machines* (Furey et al., 2000; Brown et al., 1999) e algoritmos genéticos (Zwir et al., 2002; Ooi and Tan, 2003; Deb and Reddy, 2003; Liu et al., 2005; Mitra and Banka, 2006). Em todos os projetos citados anteriormente, o objetivo é encontrar conjuntos de genes (*clusters*) que possam ser utilizados como classificadores confiáveis, com uma elevada taxa de classificação e um bom desempenho de generalização. Dessa forma, os conjuntos minerados

podem auxiliar na classificação de novos casos, facilitando o diagnóstico e o tratamento de doenças. Entretanto, em nenhum desses trabalhos, encontramos classificadores baseados em regras de alto nível, por exemplo, regras do tipo IF-THEN. Ao contrário, os classificadores obtidos são do tipo caixa-preta, onde a entrada são os dados de expressão de uma determinada amostra de células e a saída é a classe à qual essa amostra provavelmente pertence, podendo esta saída estar associada, por exemplo, a uma classe de doença. Assim, a partir de um conjunto de dados de milhares de genes chega-se a um pequeno conjunto de poucas dezenas de genes que sejam discriminantes para o problema.

Neste trabalho, o enfoque será a busca (mineração) de regras de alto nível, que não só sejam associadas a cada classe individualmente, reduzindo o problema a poucos genes por classe, mas também associando o nível de expressão gênica a cada gene que compõe a regra. Acreditamos que esse tipo de informação possa ser de grande utilidade aos especialistas que buscam entender o mecanismo por detrás de alterações nos padrões de expressão gênica associadas ao aparecimento de determinadas doenças. Para tal, elaborou-se um Algoritmo Genético para a obtenção de regras do tipo IF-THEN a partir de bases de dados de expressões gênicas. Este ambiente evolutivo foi aplicado na classificação de uma base de dados de expressões gênicas de células cancerígenas, advindas de experimentos de *microarray* (Ross et al., 2000). O principal objetivo foi a busca das relações entre os níveis de expressões gênicas de nove classes de câncer: mama, sistema nervoso central, colom, leucemia, melanoma, pulmão, ovário, renal e células reprodutivas. Como ponto de partida, utilizamos conjuntos reduzidos de genes que foram minerados a partir do trabalho anterior nessa mesma base de dados (Ooi and Tan, 2003).

2 Algoritmos Genéticos (AGs)

AGs são métodos computacionais de busca baseados nos mecanismos da evolução natural e na genética, simulando a teoria da seleção natural de Darwin (Goldberg, 1989). O AG é um algoritmo que manipula, em paralelo, um conjunto de indivíduos (população), tipicamente cadeias de símbolos de tamanho fixo, que representam cromossomos. A cada indivíduo está associada uma avaliação. O AG transforma a população corrente em uma nova população usando operações de reprodução e sobrevivência, segundo critérios baseados na função de avaliação (Koza, 1992).

2.1 Aplicações de Algoritmos Genéticos em Data Mining e em Expressão Gênica

Data Mining é um conjunto de técnicas e ferramentas aplicado para a descoberta do conhecimento em bases de dados. A tarefa de classificação é uma das várias estudadas em *data mining*. Em essência, o problema consiste em atribuir valores para os registros pertencentes a um pequeno conjunto de classes, e assim, descobrir algum relacionamento entre estes atributos. Cada registro é composto de um conjunto de atributos preditivos e um atributo objetivo (Hand, 1997; Freitas and Lavington, 1998).

O conhecimento descoberto é usualmente representado na forma de regras de predição do tipo IF-THEN. Este tipo de regra se destaca devido ao seu alto nível de entendimento e pela representação do conhecimento simbólico, contribuindo para compreensibilidade das informações descobertas. As regras descobertas podem ser construídas de acordo com vários critérios, tais como: grau de confiança da predição, taxa de acerto da classificação para amostras de classes desconhecidas, compreensibilidade, dentre outros (Fidelis et al., 2000).

Dentre os vários trabalhos que foram desenvolvidos utilizando AGs na solução de tarefas de *data mining* podemos citar (Fidelis et al., 2000; Carvalho and Freitas, 2000; Miranda. et al., 2003; Tan et al., 2003; Ishibuchi and Yamamoto, 2004; Ding et al., 2005).

Uma outra área onde os AGs estão contribuindo para a descoberta de conhecimento é a área de expressão gênica (Zwir et al., 2002; Mitra and Banka, 2006; Deb and Reddy, 2003; Ooi and Tan, 2003; Liu et al., 2005). A maioria destes projetos busca clusterizar conjuntos de genes na busca de relações entre estes genes, objetivando assim, encontrar conjuntos de genes que são classificadores confiáveis, que auxiliam na classificação de novos casos, facilitando o diagnóstico e o tratamento de tumores cancerígenos.

3 Ambiente Evolutivo

O modelo do AG empregado em nosso ambiente evolutivo foi adaptado a partir do modelo de AG proposto em (Fidelis et al., 2000). O AG em (Fidelis et al., 2000) foi elaborado com o objetivo de obter regras de classificação do tipo IF-THEN em bases de dados clínicos de pacientes. Dessa forma, as bases de dados onde o ambiente de Fidelis e colaboradores foram aplicadas eram formadas por registros que se caracterizavam por dados do paciente (idade e presença da doença em histórico familiar) e por dados relacionados a sintomas do paciente. As características que se relacionavam aos sintomas, que eram a maioria, foram todas discretizadas em: 0 - ausente, 1 - ocorrência

leve, 2 - ocorrência moderada e 3 - ocorrência severa. Nosso ambiente evolutivo, implementado na linguagem Delphi®, precisou ser adaptado para trabalhar com bases de dados de expressão gênica, onde os registros apresentam os níveis de expressão de dezenas ou centenas de genes, que são valores contínuos e com precisão variável (números reais). Para se chegar no ambiente evolutivo utilizado neste trabalho, partimos dos parâmetros propostos em (Fidelis et al., 2000) e fomos, experimentalmente, ajustando-os para o nosso ambiente. Vários aspectos foram abordados, tais como: melhores métodos de seleção e reinserção, tamanho da população, número de gerações, peso, tamanho do *tour* e precisão (número de casas após a vírgula). A seguir as principais características de nosso modelo de AG são detalhadas: codificação do indivíduo, operadores genéticos e função de avaliação.

3.1 Cromossomo ou Indivíduo

O indivíduo ou cromossomo do nosso AG é composto por n genes, onde cada gene do indivíduo está relacionado a uma condição envolvendo um atributo (um gene do *dataset*), onde n é o número de genes encontrados na base de expressão gênica. A primeira posição do indivíduo corresponde ao primeiro gene encontrado na base de dados e assim sucessivamente até que todos os genes de cada *dataset* estejam representados. O indivíduo é ilustrado na Figura 1.

$Gene_1$				...	$Gene_N$			
I_1	P_1	O_1	V_1	...	I_N	P_N	O_N	V_N

Figura 1: Cromossomo ou Indivíduo

Cada i-ésima posição do indivíduo é subdividida em quatro campos: I (índice), P (peso), O (operador) e V (valor) como ilustrado acima. Cada gene corresponde a uma condição na parte SE da regra e o indivíduo (cromossomo) a toda parte antecedente da regra. O campo I armazena o código do gene encontrado na base de dados e pode variar do valor 1 ao valor 1000. O campo P é uma variável do tipo inteira e o seu valor está compreendido entre os valores 0 (zero) e 10 (dez). É importante dizer que este campo P é o responsável pela inserção ou exclusão do gene na regra. Caso este valor seja menor do que um valor limite este gene não fará parte da regra, caso contrário o mesmo fará. Neste trabalho, partimos dos valores encontrados em (Fidelis et al., 2000) e após diversos ajustes foi utilizado como limite o valor 8 (oito). O campo O pode variar entre as operações $<$ (menor) e $\geq$ (maior ou igual). O campo V é uma variável do tipo ponto flutuante que pode variar entre o menor e o maior valor encontrados na base de expressão gênica avaliada.

3.2 Operadores Genéticos

Na seleção dos pais para *crossover* aplicamos o método do Torneio Estocástico utilizando *tour* de tamanho 3 (três). Nestes pais selecionados, aplicamos *crossover* múltiplo com dois pontos de corte, gerando dois novos filhos com taxa de *crossover* de 100%. Nestes dois filhos gerados, aplicamos o operador de mutação. Os operadores de mutação utilizados neste trabalho variam com o tipo do gene avaliado e possui taxa de mutação por gene no valor de 30%. Para chegarmos a este valor de taxa de mutação, partimos dos valores encontrados em (Fidelis et al., 2000) e fomos ajustando buscando o melhor valor para nossa aplicação. Para o gene P o novo valor é dado sorteando o incremento ou o decremento de um (1) ao valor original. Para o gene O ocorre o sorteio de um novo operador dentre os possíveis excluindo o encontrado originalmente. Neste trabalho foi utilizado apenas dois operadores ($<$ e $\geq$), levando à troca de um pelo outro quando aplica-se o operador de mutação ao gene O. A mutação do gene V é feita sorteando o incremento ou o decremento de 0,1 ao valor original. Na composição dos indivíduos que irão participar da próxima geração do AG, selecionamos os melhores pais e filhos.

3.3 Função de Avaliação ou Aptidão (FA) (Fitness Function)

A Aptidão (ou *fitness*) refere-se ao grau de contribuição de uma determinada solução candidata para a convergência do AG na busca da melhor solução dentro do espaço de busca.

Neste trabalho, a FA avalia a qualidade de cada regra (indivíduo). A FA aqui aplicada pode ser encontrada em (Lopes et al., 1997). Para o entendimento da FA aqui aplicada, alguns conceitos precisam ser elucidados. Quando aplicamos uma regra na classificação de um caso, quatro diferentes resultados podem ser observados, dependendo da classe predita pela regra e a da verdadeira regra do caso. São eles:

- *True Positive* (*tp*) - A regra prediz que o caso pertence a uma determinada classe e o mesmo pertence;
- *False Positive* (*fp*) - A regra prediz que o caso pertence a uma determinada classe mas o mesmo não pertence;
- *True Negative* (*tn*) - A regra prediz que o caso não pertence a uma determinada classe e o mesmo não pertence;
- *False Negative* (*fn*) - A regra prediz que o caso não pertence a uma determinada classe mas o mesmo pertence;

A FA utiliza dois indicadores comumente utilizados em domínios médicos, chamados de sensi-

bilidade (Se) e especificidade (Sp). Se e Sp são definidos abaixo:

$$Se = \frac{tp}{(tp + fn)} \quad (1)$$

$$Sp = \frac{tn}{(tn + fp)} \quad (2)$$

Finalmente, a FA utilizada é definida como o produto destes dois indicadores, Se e Sp, como segue abaixo:

$$FA = Se * Sp \quad (3)$$

O objetivo do trabalho é maximizar ao mesmo tempo Se e Sp e conseqüentemente o valor de FA, utilizando para isso, as equações 1, 2 e 3. Em cada execução, o nosso AG trabalha com um problema de classificação de duas classes, isto é, quando o AG está procurando por regras de uma dada classe, todas as outras classes são agrupadas em uma única classe.

3.4 Bases de dados

As bases utilizadas no nosso trabalho foram extraídas do trabalho (Ooi and Tan, 2003). Este trabalho partiu de conjuntos de genes extraídos da base NCI60 (Ross et al., 2000) composta por dados de expressão gênica advindos de experimentos de *microarray*, experimentos estes, contendo informações sobre células cancerígenas de 9 (nove) classes. São elas: mama, sistema nervoso central, cólom, leucemia, melanoma, pulmão, ovário, renal e células reprodutivas. Estas bases de dados, chamadas de B1 e B2, são formadas por conjuntos de genes preditores, contendo 13 e 12 genes respectivamente.

4 Resultados

Na obtenção destes resultados utilizamos, como parâmetros do AG, uma população inicial de 400 indivíduos e o executamos por 100 gerações. O AG foi aplicado sobre 2/3 dos registros de cada base B1 e B2 (Ooi and Tan, 2003), e o restante dos registros (1/3), foi utilizado na validação das regras obtidas pelo AG.

Como é possível observar na Tabela 1, embora os resultados de treinamento sejam bem próximos a 100% nas duas bases avaliadas, o resultado de generalização dessas regras não é tão bom, pois ao aplicarmos as mesmas sobre a terceira partição dos registros que ficaram de fora da evolução do AG, o resultado de classificação das regras cai para 80,7% de média. Acreditamos que esse desempenho se deva ao baixo número de amostras por classe que, em alguns casos chega a apenas 4 (quatro) registros por classe. Assim, realizamos várias execuções do AG na esperança de que ao obtermos uma variedade de regras com 100% de treinamento para cada classe, pelo menos uma delas tivesse uma boa capacidade de generalização (alto valor de teste).

Tabela 1: Média geral

	Média Geral	
Base	Treinamento	Teste
B1	0,996780	0,819889
B2	0,987888	0,794222

A Tabela 2 traz os melhores resultados obtidos nessa busca, apresentando as melhores regras descobertas pelo nosso AG. Para cada cada classe, nosso ambiente evolutivo foi executado 50 (cinqüenta) vezes, variando a semente randômica utilizada na geração da população inicial. A melhor regra encontrada nas 50 execuções, levando em consideração seu valor de treinamento em dois terços dos registros (e usando o menor número de genes como critério de desempate) foi selecionada como a regra preditora da classe. Cada uma destas regras foi aplicada separadamente em uma nova amostra de teste (1/3 dos registros), para avaliar o do nível de generalização de cada regra obtida em treinamento.

Tabela 2: Melhores Resultados

C	**Regra**	**Trein**	**Teste**	**Base**
1	if(Gene_46<1,8) and (Gene_289<0,5) and (Gene_306<0,3) and (Gene_783<0,1) and (Gene_865≥-1,2)	0,917	0,444	B2
2	if(Gene_11≥0,4) and (Gene_289<-0,5)	1	1	B1,B2
3	if(Gene_50<-2,3) and (Gene_194<-1,1) and (Gene_289≥-0,3)	1	1	B1
4	if(Gene_50≥-2,1) and (Gene_194<-0,7) and (Gene_366<-0,1)	1	1	B1
5	if(Gene_289≥-1,5) and (Gene_380<-0,7) and (Gene_661≥-1,2)	1	1	B2
6	if(Gene_97≥-1,4) and (Gene_242<0,3) and (Gene_828<0,1) and (Gene_839≥-0,5) and (Gene_863≥-0,3)	1	0,667	B1
7	if(Gene_97<1,4) and (Gene_194≥0,2) and (Gene_839<-0,2)	1	0,5	B1
8	if(Gene_97≥0,7) and (Gene_348<-0,8) and (Gene_863<0,7)	1	1	B1
9	if(Gene_11≥-3,6) and (Gene_177<-2)	0,974	1	B2

Para cada regra encontrada na Tabela 2 mostramos informações da sua avaliação em um conjunto de treinamento e teste, obtidos através da equação 3, além da base de dados na qual a regra foi minerada.

Das nove classes avaliadas, em cinco delas (classes 2, 3, 4, 5 e 8) foi possível atingir 100% de avaliação, tanto em treinamento quanto em teste. Na classe 9 o resultado também foi bom, pois encontramos uma regra que obteve 97,4% de avaliação em treinamento e 100% em teste. Para a classe

6, o resultado não foi tão bom, pois obtivemos 100% de avaliação em treinamento mas somente 66,7% em teste. Para as outras duas classes, 1 e 7, os resultados foram piores do que os encontrados para a classe 6. Obtivemos 91,7% em treinamento e 44,4% em teste para a classe 1 e 100% em treinamento e 50% em teste para a classe 7. Assim, consideramos que o desempenho foi muito bom em seis das nove classes, mas bem abaixo do satisfatório nas outras três.

Os valores de *fitness* ilustrados na Tabela 2 foram evoluídos levando-se em consideração cada regra separadamente. Mesmo não sendo o principal objetivo deste trabalho, uma outra forma de analisarmos este conjunto de regras é como um classificador caixa-preta, onde as regras são avaliadas como um conjunto e não separadamente. Neste enfoque, conseguimos um bom classificador, obtendo 85,25% de acertos.

A Tabela 3 ilustra os 9, num total de 61, registros da base NCI60 que não foram classificados corretamente pelo conjunto de regras. Pode-se observar que dos 9 erros encontrados 6 foram classificados corretamente pelo conjunto de regras mas também classificaram outras classes. Os conflitos encontrados foram: (i) entre as classes 1 e 5 em quatro registros, (ii) entre as classes 1 e 4 em um registro e (iii) entre as classes 4 e 9 em um registro. Em três registros nenhuma regra foi disparada. Nenhum desses casos apresentados acima representa um erro grave, por exemplo, o registro é da classe 1 e a classe que o dispara é somente da classe 2.

Tabela 3: Erros de classificações encontrados para a base NCI60

Registro NCI60	Classe	Regras Disparadas
19	C4	C4 , C9
22	C5	C5, C1
23	C5	C5, C1
24	C5	C5, C1
29	C6	Nenhuma
45	C1	Nenhuma
50	C4	C4, C1
53	C5	C5, C1
58	C7	Nenhuma

5 Conclusão e Trabalhos Futuros

Em nossos experimentos, foi possível observar que embora a obtenção de regras com alto índice de treinamento seja relativamente fácil, a qualidade dessas regras é logo diminuída em algumas classes pelo desempenho das mesmas na base de testes. Acreditamos que tal comportamento possa ser justificado pelo baixo número de amostras por classe, inerente ao problema. Para compensar essa dificuldade, procuramos executar um grande número de execuções do AG, para obtenção de um maior número de regras por classe, com alta taxa de desempenho na base de treinamento. Dessa forma, conseguimos obter excelentes regras em seis das nove classes. Entretanto, em três classes não foi possível obter regras satisfatórias. Animados com os resultados promissores desse trabalho, pretendemos dar continuidade ao mesmo com os seguintes passos: (i) analisar outras duas bases de dados extraídas de (Ooi and Tan, 2003), todas advindas de experimentos de mineração da base NCI60 (Ross et al., 2000), contendo 17 e 20 genes respectivamente; (ii) aplicar a abordagem desenvolvida neste trabalho em outras novas bases criadas através da composição das quatro bases encontradas em (Ooi and Tan, 2003).

Com as regras de alto nível obtidas, e com as que ainda serão obtidas em novos experimentos, acreditamos que será possível delimitar genes relacionados a cada classe de câncer e seus respectivos níveis de expressão. Desta forma, obteremos uma associação gene/câncer e gene/gene que esperamos que possa contribuir para o diagnóstico deste tipo de câncer limitando assim o número de genes a serem analisados na busca de novos tratamentos.

AGRADECIMENTOS

G.M.B.O. agradece ao CNPq (PQ:304639/2004-4) e à Fapemig pelo suporte.

Referências

Alberts, B., Bray, D. and Lewis, J. (1997). *Biolgia Molecular da Célula*, 3 edn, Artes Médicas.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: the Machine Learning approach*, 2 edn, MIT Press.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. and Haussler, D. (1999). Knowledge-based analysis of microarray gene expression data by using support vector machines, *Stanford University of Medicine* .

Carvalho, D. R. and Freitas, A. A. (2000). A genetic algorithm-based solution for the problem of small disjuncts, *in* Springer-Verlag (ed.), *Principles of Data Mining and Knowledge Discovery*, Vol. 1910, 4th European, Lecture Notes in Artificial Intelligence, Lyon, France, pp. 345–352.

Deb, K. and Reddy, A. R. (2003). Classification of two and multi-class cancer data reliably using multi-objective evolutionary algorithms, *KanGAL Report* .

Ding, H., Benyoucef, L. and Xie, X. (2005). A simulation-based multi-objective genetic algorithm approach for networked enterprises optimization, *Engineering Applications of Artificial Intelligence* .

Fidelis, M. V., Lopes, H. S. and Freitas, A. A. (2000). Discovery comprehensible classification rules with a genetic algorithm, *Congress on Evolutionary Computation - (CEC-2000)*, La Jolla, CA, USA, pp. 805–810.

Freeman, W. M., Walker, S. J. and Vrana, K. E. (1999). Quantitative rt-pcr: pitfalls and potentials, *Biotechniques* **26**: 112–122.

Freitas, A. A. and Lavington, S. H. (1998). *Mining Very Large Databases with Parallel Processing*, Kluwer Academic Publishers, London.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Oxford University Press* .

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*, Adison-Wesley, USA.

Hand, D. (1997). *Construction and Assessment If Classification Rules*, John Wiley and Sons, Chichester.

Ishibuchi, H. and Yamamoto, T. (2004). Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining, *Fuzzy Sets and Systems* (141): 59–88.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classifiction and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine* .

Koza, J. R. (1992). *Genetic Programming. On the Programming of Computers by Means of Natural Selection*, MIT Press, USA.

Liu, J. J., Culter, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L. and Ling, X. (2005). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Oxford University Press* **21**(11 2005): 2691–2697.

Lopes, H. S., Coutinho, M. S. and Lima, W. C. (1997). An evolutionary approach to simulate cognitive feedback learning in medical domain, *in* E. Sanchez, T. Shibata and L. A. Zadeh (eds), *Genetic Algorithms and Fuzzy Logic Systems*, World Scientific, pp. 193–207.

Miranda., C. R. S., Oliveira, G. M. B. and Santos, J. B. (2003). Algoritmos genéticos aplicados em data mining para obtenção de regras simples e precisas, *Anais do SBAI2003*, pp. 638–643.

Mitra, S. and Banka, H. (2006). Multi-objective evolutionary biclustering of gene expression data, *Pattern Recognition* .

Ooi, C. H. and Tan, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatic* **19**(1): 37–44.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics* .

Setúbal, J. C. and Meidanis, J. (1997). *Introduction to Computacional Molecular Biology*, PWS Publishing Company, Boston.

Souto, M. C. P., Lorena, A. C., Delbem, A. C. B. and de Carvalho, A. C. P. L. F. (2003). Técnicas de aprendizado de máquina para problemas de biologia molecular, Sociedade Brasileira de Computação, Sociedade Brasileira de Computação, Porto Alegre.

Tan, K. C., Yu, Q., Heng, C. M. and Lee, T. H. (2003). Evolutionary computing for knowledge dicovery in medical diagnosis, *Artificial Intelligence in Medicine* (27): 129–154.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995). Serial analysis of gene expression, *Science* **270**: 484–487.

Xu, Y., Selaru, F. M., Yin, J., Zou, T. T., Shustova, V., Mori, Y., Sato, F., Liu, T. C., Olaru, A., Wang, S., Kimos, M. C., Perry, K., Desai, K., Greenwald, B. D., Krasna, M. J., Shibata, D., Abraham, J. M. and Meltzer, S. J. (2002). Artificial neural networks and gene filtering distinguish between global gene expression profiles of barret's esophagus and esophageal cancer, *Cancer Research* .

Zwir, I., Zaliz, R. R. and Ruspini, E. H. (2002). Automated biological sequence description by genetic multiobjective generalized clustering, *New York Academy of Sciences* (980): 65–82.