Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines

Sihua Peng[a], Qianghua Xu[b], Xuefeng Bruce Ling[c], Xiaoning Peng[d], Wei Du[a], Liangbiao Chen[b,e,*]

[a] National Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, PR China
[b] College of Life Sciences, Zhejiang University, Hangzhou 310029, PR China
[c] Tularik Inc., South San Francisco, CA 94080, USA
[d] Department of Molecular Genetics, MD Anderson Cancer Center, University of Texas, Houston, TX 77030, USA
[e] Institute of Genetics and Developmental Biology, The Chinese Academy of Sciences, Beijing 100101, PR China

Received 12 September 2003; revised 23 October 2003; accepted 30 October 2003

First published online 12 November 2003

Edited by Thomas L. James

Abstract **Simultaneous multiclass classification of tumor types is essential for future clinical implementations of microarray-based cancer diagnosis. In this study, we have combined genetic algorithms (GAs) and all paired support vector machines (SVMs) for multiclass cancer identification. The predictive features have been selected through iterative SVMs/GAs, and recursive feature elimination post-processing steps, leading to a very compact cancer-related predictive gene set. Leave-one-out cross-validations yielded accuracies of 87.93% for the eight-class and 85.19% for the fourteen-class cancer classifications, outperforming the results derived from previously published methods.**

Key words: Microarray; Support vector machine; Genetic algorithm; Recursive feature elimination; Cancer

1. Introduction

Microarray technology allows the genomic-scale study of biological processes through simultaneous monitoring of the relative expression values of thousands of genes. One of the most promising applications of the technology is molecular diagnosis, by which the nature of biological samples could be identified based on their gene expression data [1–13]. In principle, molecular cancer diagnosis can be mathematically formulated as classification tasks. Most previous studies [1–3,5,7,14] in this area offer algorithmic solutions for binary classifications, e.g. tumor versus normal tissue, positive treatment effect versus no response. Because of the large number of cancer types and subtypes, it is imperative to develop multiclass tumor identification methodologies for practical cancer diagnosis purposes. The rank-based gene selection binary classification methodologies [15], however, cannot be extended to produce comparable accuracies for simultaneous multiclass classification. Therefore, some hybrid methods [8–10,13] have been developed to statistically derive confident multiclass tumor classifiers utilizing one versus all (OVA) or all paired (AP) binary support vector machine (SVM) [16] classifications. Effective feature reduction and identification of discriminant genes can lead to novel clinical reagents and be of practical interest to multiclass tumor medical diagnostic tests. Recursive feature elimination (RFE [8]), genetic algorithm (GA [7,13,17,18]) and ranking [10] approaches have been employed in order to yield a compact biologically relevant predictive gene set.

In this report, we have combined the methodologies of GA, AP/SVM, RFE to take the inherent advantages of these algorithms for array-based multiclass tumor classification. Processing the public NCI60 and GCM data sets, our algorithm enables the selection of a very small discriminant gene set and high leave-one-out cross-validation (LOOCV) accuracies, outperforming previously described methods.

2. Materials and methods

2.1. Data sets

The NCI60 data set is as previously described ([4,19], http://www-genome.wi.mit.edu/mpr/NCI60/NCI_60.expression.scfrs.txt for pre-processed, and http://genome-www.stanford.edu/sutech/download/nci60/index.html for processed respectively). A subset of NCI60 containing 58 samples belonging to eight cancer types was used in this study http://fishgenome.org/publication/pengsihua/pengetal-FEBS2003.htm. The GCM data set is as described by Ramaswamy et al. ([8], http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=61). The leukemia data set is as previously described ([2] http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi). The colon data set was as described ([1] http://microarray.princeton.edu/oncology/affydata/).

2.2. Classification strategies

Our multiclass classification algorithm consists of the following steps: a pre-filtering process to result in a set of genes differentially expressed across various cancer types; binary tumor classifications via AP/SVM classifier; a voting scheme to go from binary to multiclass classification based upon AP/SVM results; GA feature selection and multiclass classification optimization via LOOCV fitness test; RFE through AP/SVMs and LOOCV test to further eliminate the non-predictive features in the GA-derived gene set.

2.2.1. Pre-filter processing. As shown in Fig. 1, there were only a small number of genes in the NCI60 and GCM data sets that were differentially expressed across various cancer types. Prior to the machine learning procedure, features (genes) with lower standard deviations among the various tumor types were removed. A total of 1994 genes of the NCI60 data set were selected, the standard deviation

*Corresponding author. Fax: (86)-10-62551951.
E-mail address: lbchen@genetics.ac.cn (L. Chen).

Abbreviations: LOOCV, leave-one-out cross-validation; GA, genetic algorithm; SVM, support vector machine; RFE, recursive feature elimination; AP, all paired

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)01275-4

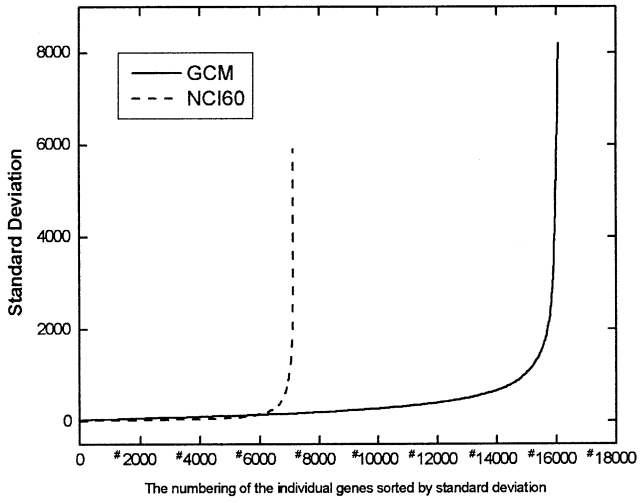

Fig. 1. Standard deviation of the expression level of each gene in the GCM and NCI60 data sets. It is apparent that only a fraction of genes (the genes around #7500 in the NCI60 data set and the genes around #16000 in the GCM data set) showed strong expression differentiation among the studied tumor types, and are suitable for subsequent classification purposes.

ranging from 78.878 to 5871.2. Two thousand genes of the GCM data set were selected, the standard deviation ranging from 682.6 to 8196.1.

2.2.2. AP/SVM and multiclass classification voting. The core algorithm of SVM in this study comes from LibSVM (http://www.csie.edu.tw/~cjlin). The AP approach was used, in which $k(k-1)/2$ classifiers were constructed to classify two different tumor classes (k is the number of tumor types). In the SVM procedure, three different types of kernel functions (linear, polynomial, and radial basis) are included in the algorithm. We found that polynomial kernels yielded better results than Gauss and sigmoid kernels. Integers between 1 and 8 were tested for the optimal value for the power of the polynomials. Our results showed that the best performance was obtained when the power of the polynomial was 4.

Going from binary to multiclass tumor classification, a voting strategy was used: each binary classification is considered to be a vote where votes can be cast for all testing samples; and at the end, a sample is assigned to a particular tumor class with the maximum number of binary classification votes. When two classes have identical votes, the one with the smaller index is selected.

2.2.3. LOOCV and GA. LOOCV is as previously described [20]. LOOCV of the multiclass classification results served as the GA's fitness test. Our source code was developed by modifying from Ooi et al. [13], and can be downloaded from http://fishgenome.org/publication/pengsihua/pengetalFEBS2003.htm. Two selection methods are employed to select the population individuals for the mating pool: (1) stochastic universal sampling (SUS) and (2) roulette wheel selection (RWS). Uniform and one-point cross-overs exchange subparts of the two chromosomes. Our results showed that the RWS strategy is better than the SUS strategy as a selection operation and the uniform strategy is better than one-point cross-over as a cross-over operation. The computation was difficult to converge when the crossover probability P_c was set to be less than 0.8. When P_c ranges from 0.98 to 1.0, our experiments showed that relatively better results were achieved. To optimize mutational probability P_m, we tried more than 10 P_m values from 0.0005 to 0.02 and found: when $P_m > 0.1$, the computation did not converge, the fitness oscillated along a lower value; when $P_m < 0.001$, the fitness did not improve unless large generation of GA was set; optimal P_m values were determined to be between 0.004 and 0.006. Generation in GA was initially set to 100 000 and fine-tuned to finalize for each classification process gauged by LOOCV rate: colon tumor binary classification, 31 527; acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) binary classification, 1219; NCI60 multiclass classification, 6729; GCM, 12 765. We tried different population sizes and found that values between 6 and 40 yielded the best results. We set the GA population size as 12 for binary classification (colon tumor and AML/ALL) and 30 for multiclass classification (NCI60 and GCM). GA chromosome size was set between 36 and 40. Since we employed a RFE post-processing step after the GA/AP-SVM process, the final selected feature number was smaller than the GA chromosome size.

2.2.4. RFE post-processing. Our AP/SVM RFE is adapted from the previous OVA/SVM RFE by Ramaswamy et al. [8]. Each AP/SVM classifier is first trained with all GA-derived genes, then one feature is removed, and each classifier is retrained with the smaller gene set. If LOOCV of the multiclass classification demonstrates that the accuracy remains the same or improves, then this gene feature is permanently eliminated from the predictive gene set. This procedure is repeated iteratively to derive the final predictive gene set.

2.3. Microarray data clustering and expression pattern visualization

Gene Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) and Java TreeView (http://genetics.stanford.edu/~alok/TreeView) tools were used as instructed.

3. Results and discussion

Table 1 summarizes our algorithm GA/SVM parameters, feature gene selection, and the LOOCV results. The best fitness we obtained from binary classifications was 100% for subtyping of leukemia (AML and ALL) with six genes, and 93.55% for colon tumor classification with as few as 12 genes.

Table 1
The parameters and outcome of the GA/SVM in the four data sets

Data set		GA	SVM	Predictive Gene Set	
				Number of features	LOOCV (%)
Leukemia	Uniform	$P_c = 1$ $P_m = 0.005$	Poly	6	100.0
	RWS	Popsize = 12 $n = 1219$	Degree = 4		
Colon	Uniform	$P_c = 1$ $P_m = 0.006$	Poly	12	93.55
	RWS	Popsize = 30 $n = 31\,527$	Degree = 4		
NCI60	Uniform	$P_c = 1$ $P_m = 0.005$	Poly	27	87.93
	RWS	Popsize = 30 $n = 6729$	Degree = 4		
GCM	Uniform	$P_c = 0.98$ $P_m = 0.002$	Poly	26	85.19
	RWS	Popsize = 30 $n = 12\,765$	Degree = 4		

Our results were either comparable or superior to those previously reported [2,4,8,13].

As for multiclass tumor classification (Tables 1 and 2), the LOOCV results showed 87.93% for the NCI60 data set, and 85.19% for the GCM data set. When compared with previously described algorithms, our methods yielded obvious improvements: two-dimensional hierarchical clustering achieved 81% LOOCV for the NCI60 data set; OVA/SVM achieved LOOCV of 78% by Ramaswamy et al. [8] and 81.25% by Yeang et al. [9] for the GCM data set; OVA/KNN (K nearest neighbor) achieved 72.92% LOOCV [9] for the GCM data set; GA/MLHD achieved LOOCV of 85.37% for the NCI60 data set and 79.33% for the GCM data set.

Our results demonstrate that the combination of the GA algorithm with SVM bestows many characteristics beneficial to microarray data analysis. GA differs substantially from other traditional search and optimization methods; they search a population of points in parallel (one generation),

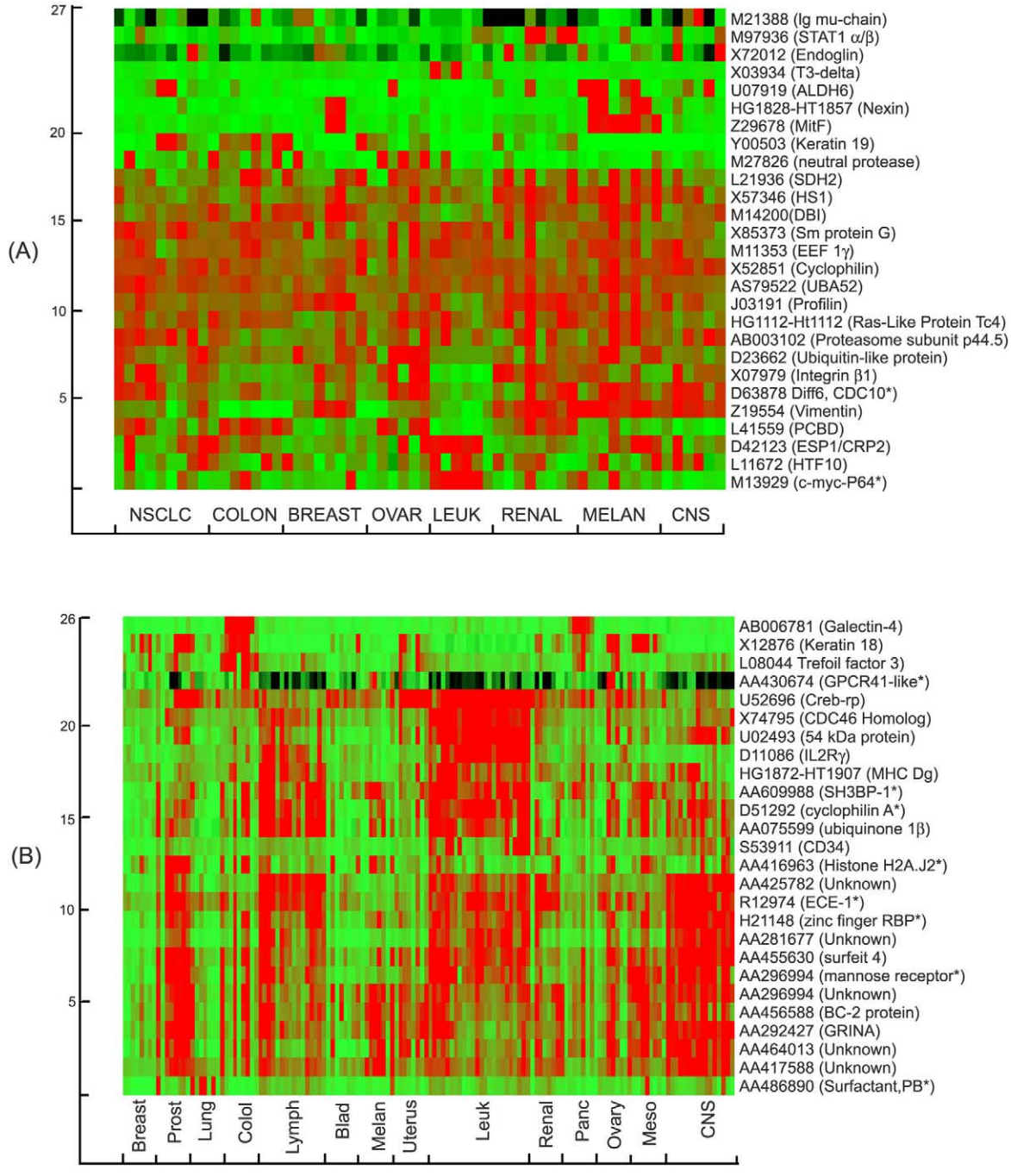

Fig. 2. Expression profiles of the predictor genes in the NCI60 data set (A), and the GCM data set (B). The *x*-axis denotes the tumor types. The accession numbers and brief descriptions of the predictor genes are shown along the *y*-axis. An asterisk by the gene description indicates the description was obtained through sequence homology search in the NCBI database. Red colored small squares represent up-regulated events, with the intensity of the redness indicating the degree of up-regulation. Green squares indicate unchanged expression levels and the black color represents down-regulated events, with the intensity of darkness reflecting the degree of down-regulation. In the figure, NSCLC denotes non-small-cell lung carcinoma; OVAR, ovary; LEUK, leukemia; MELAN, melanoma; and CNS, central nervous system.

Table 2
Results comparison of GA/SVM with some other algorithms

Classification Method	NCI60 data set		GCM data set		Reference
	LOOCV (%)	Number of features (genes)	LOOCV (%)	Number of features (genes)	
Hierarchical clustering	81	6831	–	–	[4]
OVA/SVM	–	–	78	16 063	[8]
OVA/SVM	–	–	81.25	16 063	[9]
OVA/KNN	–	–	72.92	16 063	[9]
GA/MLHD	85.37	13	79.33	32	[10]
GA/SVM/RFE	87.93	27	85.19	26	This study

rather than searching for a better result from point to point, e.g. simulated annealing. GA does not need derivative information or other auxiliary knowledge; only the fitness levels determine the directions of the search. SVM can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients), and is robust with noise. SVM also has the ability to avoid overfitting, which imposes an essential advantage over other methods. The number of support vectors selected by the learning algorithm is usually small, even with a large training data set. This characteristic would be essential even if the whole genome chip is used in the near future.

In this paper, we address the problem of selection of a compact subset of genes for simultaneous multiclass tumor classification. As shown in Table 2, our method significantly eliminated gene redundancy and yielded a more compact and unique gene subset: 27 out of 7129 genes in the NCI data set and 26 out of 16 063 genes in the GCM data set were consistently selected. Resembling the situation previously described [13], our gene set barely overlapped with those selected by other algorithms. Few selected genes matched the top 50 genes reported by Golub et al. [2]. Interestingly, although GA was employed in both studies, the two feature gene sets selected by our GA/AP SVM/RFE and the previously described GA/MLHD [13] shared only three common feature genes (D51292-cyclophilins, X12876-keratin 18, and AA416963-H2A.J2) from the GCM data set and none from the NCI60 data set. Detailed descriptions of the selected feature genes and their non-overlapping relationships with those previously published are shown in Supplementary materials (Tables 1 and 2 in http://fishgenome.org/publication/pengsihua/pengetalFEBS2003.htm).

In the NCI60 data set (Fig. 2A), Z29678 (MitF), HG1828 (nexin) and U07919 (ALDH6) were strongly up-regulated in melanoma, but very rarely up-regulated in other tumor types. It is known that MitF is crucial for the survival and proliferation of melanocyte and melanoma cells [21], and the involvement of nexin [22] and ALDH [23] in metastasis of other types of tumors was reported. X07979 (integrin β1) and the Z19554 (vimentin) gene were not expressed in the colon cancer and leukemia samples but were strongly up-regulated in other types of tumors. M21388 (Ig μ chain) and X07979 (integrin β1) were identified showing inverse regulation patterns by this algorithm in Colon, Renal and CNS tumor categories.

In the GCM data set (Fig. 2B), the correlation between the patterns of gene regulation and the tumor types was more apparent. The selected genes can be easily divided into a few groups according to their expression across the 14 tumor types. For example, bulks of genes (from AA425782 down to AA415788) are strongly up-regulated in CNS, leukemia, lymphoma and prostate cancer types, but were not or only sporadically induced in the others. AB006781 (galectin-4) was only up-regulated in colorectal and pancreas cancers, while expression of AA486890 (surfactant) was more restricted to lung cancers. A negatively regulated gene, AA430674, showed neat opposite expression patterns to the gene R12974 (ECE-1, endothelin converting enzyme 1), as well as the gene AA425782 (unknown gene), in various tumor types. AA430674 appears to be a putative G protein-coupled receptor GPCR41-like protein based upon homology search to NCBI database.

As revealed by Fig. 2, some selected genes did not demonstrate uniform expression patterns within individual samples of the same tumor type. These results suggest newer unappreciated taxonomies or reflect contributions from contaminating non-neoplastic cells, reinforcing notions brought up by previous observations [8].

In conclusion, we have combined GA, SVM, and RFE for array-based multiclass cancer classification. In comparison with previously described methodologies, our algorithms achieved higher accuracies, and derived a compact set of cancer-relevant predictive genes.

Acknowledgements: This work was supported by the National High Technology Research and Development Program of China (863 Program) No. 2002AA234011 to L.B.C. We are thankful to Tularik scientists Drs. Tim Hoey, Gene Culter and Jane Liu for critical discussions, and to the authors of C and Matlab code who made their source code available through the Internet. Our implementation grew from the source codes written by C.J. Lin et al., J.S. Ma et al. and C.H. Ooi et al. We are also thankful to Sheng Zhu (Institute of Genetics and Developmental Biology, CAS, China) for his outstanding IT support. Our source code and the supplementary materials are available at http://fishgenome.org/publication/pengsihua/pengetalFEBS2003.htm.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Proc. Natl. Acad. Sci. USA 96, 6745–6750.
[2] Golub, T.R. et al. (1999) Science 286, 531–537.
[3] Bittner, M. et al. (2000) Nature 406, 536–540.
[4] Ross, D.T. et al. (2000) Nat. Genet. 24, 227–235.
[5] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) Bioinformatics 16, 906–914.
[6] Khan, J. et al. (2001) Nat. Med. 7, 673–679.
[7] Li, L., Darden, T.A., Weinberg, C.R., Levine, A.J. and Pedersen, L.G. (2001) Comb. Chem. High Throughput Screen. 4, 727–739.
[8] Ramaswamy, S. et al. (2001) Proc. Natl. Acad. Sci. USA 98, 15149–15154.
[9] Yeang, C.H. et al. (2001) Bioinformatics 17 (Suppl. 1), S316–S322.
[10] Su, A.I. et al. (2001) Cancer Biol. 61, 7388–7393.

[11] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Machine Learn. 46, 389–422.
[12] Lu, Y. and Han, J.W. (2003) Inform. Syst. 28, 243–268.
[13] Ooi, C.H. and Tan, P. (2003) Bioinformatics 19, 37–44.
[14] Alizadeh, A. et al. (1999) Cold Spring Harbor Symp. Quant. Biol. 64, 71–78.
[15] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) J. Am. Statist. Assoc. 97, 77–87.
[16] Vapnik, V.N. (1998) Statistical learning theory, Wiley, New York.
[17] Holland, J.H. (1975) Adaptation in Nature and Artificial Systems, University of Michigan Press, Ann Arbor, MI.
[18] Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York.
[19] Staunton, J.E. et al. (2001) Proc. Natl. Acad. Sci. USA 98, 10787–10792.
[20] Rivals, I. and Personnaz, L. (1999) Neural Comput. 11, 863–870.
[21] Widlund, H.R. and Fisher, D.E. (2003) Oncogene 22, 3035–3041.
[22] Buchholz, M., Biebl, A., Neebetae, A., Wagner, M., Iwamura, T., Leder, G., Adler, G. and Gress, T.M. (2003) Cancer Res. 63, 4945–4951.
[23] Sreerama, L. and Sladek, N.E. (1997) Clin. Cancer Res. 3, 1901–1914.