An Integrated Approach of Support Vector Machine and Variable Neighborhood Search for Discovering Combinational Gene Signatures in Predicting Chemo-response of Osteosarcoma

K.Y Chan, H.L. Zhu, M.E. Aydin, C.C. Lau and H.Q. Wang

***Abstract*— Analyzing microarray data could help discover significant cancer genes and their mutual interactions, which can be used to generate hypothesis for the identification and validation of genetic biomarkers. However, the commonly used statistical significance analysis can only provide information of individual genes, thus neglecting influence of the mutual interactions. Therefore, methods aiming at discovering combinational gene signatures are highly valuable. In this paper, an integrated approach of support vector machine (SVM) and variable neighborhood search (VNS) is introduced in searching the gene signatures for predicting histologic response of chemotherapy on osteosarcoma patients. Cross validation results show that this method outperforms other existing algorithms. Further validation with the test dataset shows that only one of the fourteen samples is misclassified. The high testing accuracy further suggests that the proposed method has capability of extracting the stable discriminative signatures from microarray data.**

***Index Terms*—support vector machine, variable neighborhood search, osteosarcoma, gene signatures.**

I. Introduction

Microarray chip can simultaneously interrogate thousands of genes, which provides an extremely powerful tool for genomic studies of cancer. A few key genes (typically involving oncogenes and tumor suppressor genes), when mutated, will cause dysregulation of the transcription and translation of other genes through complicated signaling pathways to initiate oncogenesis, and ultimately leading to derangement of the cellular phenotype and the clinical manifestations of cancer. Analyzing microarray data could help discover some significant cancer genes and their mutual interactions, which can be used to generate hypothesis for the identification and validation of genetic biomarkers for diagnostic and therapeutic purposes [4, 9]. This approach has been proven to be an effective way to reveal the biological mechanism of gene regulation and pathways.

Significance based methods (e.g. T-test, Confidence intervals, etc.) [5], which aim at finding statistically significant genes in differentiating various patient groups, have been extensively utilized. However, the philosophy of these methods is to evaluate the significance on each individual gene, thus neglecting the intrinsic interactions among genes. Therefore, methods to assess the function of gene combinations in regulating tumor patterns are highly desired. Supervised classification is the most effective machine learning method to map the input space (with multiple predictor genes) to the output space (with labeled conditions). Cross-validation accuracy is generally used as a fair indicator of the performance of a classifier. The best cross-validation accuracy of a well-designed classifier thus, can implicate the relevance of the selected genes. Utilizing genuine gene markers, which potentially account for specific phenotypes, should classify samples accurately into different groups in a robust manner. Common learning algorithms include neural network [2, 13], k-nearest neighbor [15], decision tree, multi-layer perceptron [14], self-organizing maps [9], hierarchical clustering [6], graph theoretic approaches [11], and support vector machine (SVM) [8, 20, 25]. Among all of them, SVM has been proven to have the best capability in controlling the tradeoff between empirical risk and model complexity to achieve good prediction [1, 15, 22, 23]. Nevertheless, exhaustively evaluating all the combinations by supervised classification within the whole gene set will lead to a NP-complete problem, which eventually becomes computationally impossible.

In this study, an integrated approach of support vector machine (SVM) with a variable neighborhood search (VNS) algorithm, namely SVM-VNS, is introduced to discover the best gene signatures of drug response monitoring and predicting for osteosarcoma patients. The rationale behind the use of VNS is its high efficiency in searching a tremendous solution space [21]. SVM-VNS incorporates gene features and SVM parameters in one common solution code. It conducts a local search with a strategic adaptation of neighborhood according to SVM estimation that makes this innovative approach outperforming other alternatives.

K.Y. Chan, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong; e-mail: richanky@inet.polyu.edu.hk.

H.L. Zhu and H.Q. Wang, Research Institute of Innovative Products and Technologies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, e-mail: Zhu.Hailong@inet.polyu.edu.hk.

M.E. Aydin, Department of Computing and Information Systems, University of Bedfordshire, Luton, United Kingdom.

C.C. Lau, Texas Children's Cancer Center, Department of Pediatrics, Baylor College of Medicine, Houston, Texas, United States

II. Problem Formulation and Solution Representation

A. Problem Formulation

Let a gene microarray dataset **D** be $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, where $\mathbf{x}_i \in \Re^m$ is the gene expression level of the i-th patient, $y_i \in \{-1,1\}$ is the condition label for binary classification problem, and m is number of gene features. The dataset after performing gene selection is defined as $\{(\ell(\mathbf{x}_i), y_i)\}_{i=1}^{l} = \ell(\mathbf{D}) \subset \mathbf{D}$ with $\ell(\mathbf{x}_i) \in \Re^{m'}$, where function ℓ selects m' ($\le m$) gene features among all the m gene features from the gene expression data set **D**. For a new sample $\mathbf{x}$, the decision function of a SVM classifier with radial-basis-function (RBF) kernel can then be defined based on the selected gene subset:

$$f(\mathbf{x}, \mathbf{D}, \ell, \sigma, C) = \mathrm{sgn}\Big(\sum_{\text{support vectors}} y_i a_i K(\ell(\mathbf{x}_i), \ell(\mathbf{x}))\Big) \quad (1)$$

where σ is the width parameter of the RBF kernel and C is the regularization parameter, a_i is solved by optimizing a quadratic function

$$W(\mathbf{a}) = \min\left(\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l}[a_i a_j y_i y_j \cdot K(\ell(\mathbf{x}_i), \ell(\mathbf{x}_j))]\right) \quad (2)$$

subject to $0 \le a_i \le C$. The support vectors are only corresponding to those items with $a_i > 0$. To develop a robust SVM model based on the training set, the leave-one-out cross-validation (LOOCV) was applied to optimize the model parameters (σ and C). In LOOCV, one sample is leaved out as testing sample, and the remained $l-1$ samples are used as training data. Let $\overline{\mathbf{D}}_k$ represent the training set $\{(\mathbf{x}_i, y_i), i = 1, \cdots k-1, k+1, \cdots, l\}$, then the overall accuracy is calculated by:

$$J(\mathbf{D}, \ell, \sigma, C) = \frac{1}{2l}\sum_{k=1}^{l}\left|f(\mathbf{x}_k, \overline{\mathbf{D}}_k, \sigma, C, \ell) + y_k\right| \quad (3)$$

Now the problems of gene feature selection and SVM parameter optimization are integrated to optimizing the above objective function.

B. Solution Representation

Solution of the above function is represented in both binary and real codes, which consists of two parts: a) binary coded representation [3, 13] for the gene features selection, and b) real coded representation for the SVM parameter optimization. The representation of the solution is illustrated in Figure 1.

As shown in the left side of Figure 1, binary coded representation is composed of a m-bit binary string with each bit is corresponding to one gene. In the binary string, there are m' 1-elements and the rest are 0-elements. A bit with 1-element means that the corresponding gene is selected, otherwise that gene is ignored. For instance, a solution [0,1,0,1,0,0] with $m' = 2$ and $m = 6$ represents the 2nd and 4th genes are selected. Thus the number of possible gene subsets can be calculated as

$$n = \binom{m}{m'} \quad (4)$$

As illustrated in the right side of Figure 1, real code is adopted for representing the two SVM parameters, the kernel width parameter σ and the regularization parameter C.

In general, the number of the original genes contained in microarray data is very large and include lots of noisy genes, which not only impair the efficiency but also the effectiveness of any searching algorithms. Therefore, a pre-screening procedure to filter out those noisy genes is still necessary.

III. Algorithm of SVM-VNS

Variable neighborhood search (VNS) is a well-known local search algorithm [18, 21], and has been successfully applied to solve hard combinational optimization problems [7, 10]. In this paper, an integrated approach of support vector machine (SVM) and VNS algorithm, that can effectively solve the problem of simultaneously optimizing gene subset and the classifier, is introduced, namely SVM-VNS.

A. Neighborhood Structures (NSs)

The neighborhood structures NSs in the solution representation, which determines neighboring solutions to move to, is a key element of VNS in which two NSs are used:

- 'MutationBin' is a function used to explore solutions of the binary representation by exchanging the entries of a 0- and 1- elements. For instance, suppose that the 2nd bit with entry 1-element and 5th bit with entry 0-element of the solution [0,1,0,1,0,0] are selected to be exchanged. After applying MutationBin, the new solution will be [0,0,0,1,1,0]. Obviously, the elements in 2nd and 5th bits were exchanged.
- 'MutationReal' is a fine-tuning function that implies small shake on a randomly choice of classifier parameters in the real coded representation of the solution. The MutationReal function is defined as the following *Shake Function*:

$$shake(p) = p + \omega \quad (5)$$

where p represents the randomly chosen parameter, and ω is randomly generated within the range $0.1 \times (p_{\max} - p_{\min})$, representing 0.1 times scale of the parameter space of the classifier.

B. Procedures of SVM- VNS

As illustrated in step 1 of the following pseudo-code, SVM-VNS starts with a randomly selected initial solution, $[\ell, \sigma, C] \to x \in S$, where S is the whole search space, and manipulates the solutions via steps 2(a) and 2(b), where two main functions of *Shake* and *Local Search* for intensification and exploration are employed in neighborhood search. In step 2(a), *Shake Function* generates and modifies the solutions regardless of the quality of solution so as to initialize a fresh search in a local neighborhood or to switch to another neighborhood. Then step 2(b) carries out the major intensive search by *Local Search Function*, which seeks for an improved solution within the chosen local neighborhood. The outcome of *Local Search Function* is evaluated cross-validation with SVM to determine whether or not to accept it as the solution for further search. This process is iterated until the termination condition is met.

```
1. Initialization: initialize a solution $[\ell, \sigma, C] \rightarrow x \in S$
2. Repeat the following steps until the stopping condition is met:
  (a) Perform Shake Function: $x \rightarrow x'$
  (b) Perform Local Search Function: $x' \rightarrow x''$
  (c) Improve or not: if $x''$ is better than x, do $x'' \rightarrow x$
```

Shake Function and *Local Search Function* need to be chosen so as to achieve an efficient SVM-VNS. The two NSs discussed in Section III.A are used for Shake and Local Search Functions to obtain neighborhood changes for SVM-VNS.

Shake Function works to switch to another region of the search space so as to carry out a new local search, as the purpose of *Shake Function* is to diversify the exploration. Here *Shake Function* does not work in the binary coded representation, but is designed to conduct a random move within the real coded representation. For example, the given solution x^* is operated with the *Shake Function* to obtain *x'*=MutationReal(x*).

Local Search Function is developed as a simple hill-climbing algorithm based on both aforementioned NSs. As indicated in the following pseudo-code, the NSs are used complementary to each other that the functioning NS keeps iterating as long as better moves are resulted. It switches to the other NS once the result produced is not better and the algorithm stops if the number of moves, *n*, meets a predefined number, $n_{\max}$. The change of NS is organized with a binary integer variable, $\gamma \in (0, 1)$, in which the value of γ is changed by using an absolute function denoted by $|\cdot|$ norm at the second part of step 3(b) of the pseudo-code. The procedures of *Local Search Function* are as follows:

```
1. Get initial solution, $x' \in S$
2. Set $n \leftarrow 0$ and $\gamma \leftarrow 1$
3. While $n < n_{\max}$ do
  (a) if $(\gamma = 1)$ then $x' \leftarrow$ MutationBin(x'); else if $(\gamma = 0)$ then $x' \leftarrow$ MutationReal(x')
  (b) Set if $J(x') < J(x'')$ then $x' \leftarrow x''$; else $\gamma \leftarrow |l - 1|$
  (c) $n \leftarrow n+1$
where $J(x)$ is defined by (3).
```

IV. Data Description

It was reported that osteosarcoma accounts for 60 percent of malignant bone tumors diagnosed in the first two decades of life [16]. The DNA microarray data [17] were measured on tumor tissues sampled from patients before and after chemotherapy treatments. The patients are followed up to record clinical outcome such as response, recurrence and survival etc. A total of 34 samples (14 initial biopsies and 20 definitive surgery specimens) were included in this study, which were obtained from 28 individual patients with 18 males and 10 females. Six patients contributed two samples each, both initial biopsies and definitive surgery specimens, whereas the remaining 22 patients contributed one sample each, either initial biopsy or a definitive surgery specimen. The initial biopsy samples were obtained at the time of diagnosis before the initiation of preoperative chemotherapy. The definitive surgery samples were collected after the completion of preoperative chemotherapy. The drug responses are centrally reviewed by one pathologist after definitive surgery. Good response is defined as more than 90 percent necrosis in tumor, and poor response with less than 90 percent necrosis.

Raw quantification output of all array experiments were preprocessed and filtered by removing spots with low signal intensity and low sample variance ($P > 0.01$) as well as those that were missing in >50% of the experiments. A total of 1,934 genes remained after pre-processing and filtering. Intensities were then normalized by intensity dependent local weighted regression method. After normalization, intensity ratios were log transformed before further analysis.

There were some missing data after filtering. In this study, we simply imputed those missing values by the mean value over the training set in each cycle of cross-validation test. This approach ensures that the testing data are entirely independent to the training process to exclude any possibility of overestimation.

V. Results and Discussion

With SVM-VNS, the model was firstly trained and optimized on the 20 definitive surgery samples to classify good responders and poor responders. The best gene combination can be determined according to the cross-validation accuracy. The gene features and SVM model will then be validated with the 14 initial biopsy samples. The results can reveal the relevance of the proposed method.

Classification on definitive surgery samples. Two-sample t-test was performed on the 20 definitive surgery samples to pre-screen those noisy genes. 200 of most significant genes were kept, and others are ignored. SVM-VNS trials of using gene numbers of 2, 3, 4, 5, 10 and 20, were then performed. For each trial, the training process was repeated with 50 runs. Mean values of cross-validation accuracies were listed in Table 1.

As shown in Table 1, the accuracy initially increased when more genes were involved in the feature subset. This observation accords the logic that multiple genes cooperate together to regulate chemo-response. Larger gene set should provide more comprehensive information for classification. An interesting finding is that the accuracy declines as number of genes increased to 10 and 20. This result seems to conflict with previous reasoning. However, the fact is, when a gene is selected as feature, its expression data contributes not only information but also noise [9]. The information of a newly added gene feature could be redundant to previously selected genes, while its noise could deteriorate the classification accuracy. From this table, it is clear that 5 gene signatures can stably support the whole feature space to describe the mapping relationship between genetic expression and chemo-response.

Results of various supervised classification algorithms (compound covariate predictor CCP, linear discriminant analysis LDA, 1-neraest neighbor 1-NN, 3-nearest neighbor 3-NN, nearest centroid NC and SVM) obtained from [17] were compared with SVM-VNS. For all algorithms, LOOCV was used to evaluate the performance of classification. The SVM-VNS achieved outstanding performance (92% classification accuracy) comparing to the other six algorithms that can only achieve 60-70% classification accuracies. This result indicate that SVM-VNS is a powerful technique in achieving accurate classification through optimizing feature subset and the parameters of the SVM classifier.

Combinational gene signatures. Among the total 50 runs in the trial of 5 gene features, four subsets of gene features with 100 percent cross-validation accuracy were selected as signatures, as listed in Table 2. In this table, the gene *Enah/Vasp-like* (EVL, also known as RNB6) appears in all the subsets. It was reported that RNB6 has been identified as a commonly down-regulated gene biomarker in various types of cancers [12, 19]. Another gene *Cell division cycle 23, yeast, homolog* (known as CDC23), is a large multisubunit ubiquitin-protein ligase required for the ubiquitination and degradation of G1 and mitotic checkpoint regulators. When overexpressed, CDC23 could lead to abnormal levels of anaphase-promoting complex (APC/C) targets. [24]. Some other genes, such as *Early growth response 1* and *C1q and tumor necrosis factor related protein 2, etc.*, also have relationship with oncogenesis or tumor development, but the details would not be discussed here.

There is no available knowledge to explain the cooperative relationship among the genes in each subset. Therefore we cannot verify the validity of the selected genes as biomarkers. Nevertheless, the results can be used as a hypothesis for further investigations. Performing real-time RT-PCR can validate these genes markers. More molecular studies should be pursued to investigate the biological mechanism of these genes' involvement in determining drug response and chemoresistance.

Prediction of chemo-response on initial biopsy samples. All of the four gene subsets were then respectively applied to predict the chemo-response of the 14 initial biopsy samples for further validation purpose. For all the gene subsets, the corresponding SVM classifier misclassified only one sample among the 14 samples, with a correct classification rate of 92.9%. As shown in Table 3, they all misclassified Tumor ID 410 as a poor responder but clinically the patient was considered to be a good responder. This finding is consistent with the results of the original study in [17], where it was pointed out that this patient initially presented with localized disease but eventually developed recurrence in the lungs 25 months after completion of therapy, suggesting that there were resistant cells present in the initial biopsy, which might have metastasized to the lungs before definitive surgery and subsequently gave rise to the recurrent tumor. These results further indicate that the gene expression signature of the resistant cells in the definitive surgery samples was already present in the initial biopsy samples at the time of diagnosis.

VI. Conclusions

In this paper, we have proposed an algorithm SVM-VNS, which is an integrated approach of variable neighborhood search and support vector machine aiming at selecting a compact gene subset and simultaneously optimizing model parameters. Applying this algorithm on osteosarcoma microarray data resulted in 92 percent of cross-validation accuracy on the training dataset outperforming other commonly used algorithms. Meanwhile, four subsets of combinational gene signatures were discovered to consistently discriminating the good responders and poor responders both on training set and testing set. Some of them are reported to have close relationship with oncogenesis and tumor development. Further validation using independent methods of measurement shall be pursued to investigate the cooperative mechanism among genes in each subset. Prediction on chemo-response using an independent dataset of initial biopsy samples achieved 92.9 percent of accuracy, reflecting that the proposed SVM-VNS algorithm is promising in selecting robust gene signatures. The results of SVM-VNS suggest that it is a very useful tool to generate hypothesis for the biomarker identification and validation.

ACKNOWLEDGEMENTS

This study is supported by the project of "clinical decision support system of cancer diagnosis and treatment" (project number: 1-BB56), funded by the Niche Area Funding of the Hong Kong Polytechnic University.

References

[1] B. Boser, I. Guyon and V. Vapnik, An training algorithm for optimal margin classifiers, *Proc. 5th Annual Workshop on Comp. Learning Theory*, 144-152, 1992.

[2] F.Z. Brill, D.E. Brown and W.N. Martin, Fast genetic selection of features for neural network classifiers, *IEEE Trans. on Neural Networks*, 3(2):324-328, 1992.

[3] R.E. Caballero and P.A. Estevez, A niching genetic algorithm for selecting features for neural network classifiers, *Proc. 8th Conf. of Arti. Neu. Net.*, 1:311-316, 1998.

[4] M. Daly and R. Ozol, The search for predictive patterns in ovarian cancer: proteomics meets bioinformatics, *Cancer Cell*, 111-112, 2002.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *Jour. Amer. Stat. Ass.*, 97(457):77-87, 2002.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, Cluster analysis and display of genome-wide expression patterns, *Proc. Nat. Acad. Sci.*, 95(14): 863–14, 1998.

[7] K. Fleszar and K. S. Hindi, New heuristics for one-dimensional bin-packing, *Comp. Oper. Res.*, 29:821-839, 2002.

[8] T.S. Furer, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16(10):906-914, 2000.

[9] T.R. Golub, D.K. Slonim, P. Tamayo, C. Hurd, M. GassenBeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Blomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring, *Science*, 286:531-537, 1999.

[10] P. Handsen, N. Mladenovic and U. Dragan Variable neighborhood search for the maximum clique, *Discrete App. Math.*, 145:117–125, 2004.

[11] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach and R. Shamir, An algorithm for clustering cDNA fingerprints, *Genomics*, 66(3):249-256, 2000.

[12] S. Hasegawa et al., Genome-wide analysis of gene expression in Intestinal-Type Gastric cancers using a complementary DNA microarray representing 23,040 genes, *Cancer Research*, 62:7012-7017, 2002.

[13] J.H. Hong and S.B. Cho, Efficient huge-scale feature selection with speciated genetic algorithm, *Pat. Recog. Lett.*, Vol. 27, pp. 143-150, 2006.

[14] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med.*, 7(6): 673-679, 2001.

[15] L. Li, C.R. Weinberg, T.A. Darden and L.G. Pedersen, Gene selection for sample classification based on gene expression data: study of

sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics*, 17(12):1131–1142, 2001.

[16] M.P. Link, M.C. Gebhardt and P.A. Meyers, *Principles and Practice of Pediatric Oncology*, 1051-1089, 2002.

[17] T.K. Man, M. Chintagumpala, J. Visvanathan, J. Shen, L. Perlaky, J. Hicks, M. Johnson, N. Davino, J. Murray, L. Helman, W. Meyer, T. Triche, K.K. Wong and C.C. Lau, Expression profiles of osteosarcoma that can predict response to chemotherapy, *Cancer Research*, 65(18):8142-8150, 2005.

[18] N. Mladenovic and P. Hansen, Variable neighborhood search. *Comp. Oper. Res.*, 24:1097–1100, 1997.

[19] J. Okutsu et al. Prediction of Chemosensitivity for patients with Acute Myeloid Leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis, *Molecular Cancer Therapeutics*, 1:1035-1042, October 2002.

[20] S. Peng, Q. Xu, X.B. Ling. X. Peng, W. Du and L. Chen, Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, *FEBS Letters*, 555:358-362, 2003.

[21] M. Sevkli and M.E. Aydin 2007, Parallel variable neighborhood search algorithms for job shop scheduling problems, *IMA Jour. of Man. Math.*, 2007 (in print).

[22] V.N. Vapnik, *Statistical Learning Theory*, Wiley Interscience, 1998.

[23] V. Vapnik and O. Chapelle, Bounds on error expectation for support vector machines, *Neural Computation*, 12:2013-2036, 2000.

[24] Wang Q, Moyret-Lalle C, Couzon F, 2003, Alterrations of anaphase-promoting complex genes in human colon cancer cells, *Oncogene*, 22(10): 1486-1490, 2003.

[25] Y.D. Zhao, C. Pinilla, D. Valmori, R. Martin and R. Simon, Application of support vector machines for T-cell epitopes prediction, *Bioinformatics*, 19(15): 1978-1984, 2003.

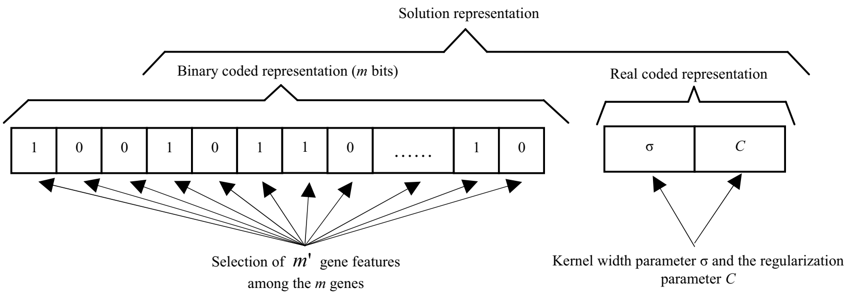
Solution representation
Binary coded representation (*m* bits)
Real coded representation
1
0
0
1
0
1
1
0
……
1
0
σ
C
Selection of m' gene features among the *m* genes
Kernel width parameter σ and the regularization parameter *C*

Figure 1. Solution representation

Table 1. Mean values of cross-validation accuracies (over 50 runs)

Number of gene features used	**Accuracy of LOOCV**
2	85%
3	85%
4	87%
5	92%
10	91%
20	90%

Table 2 Subset of combinational gene signatures

Subset	**Gene 1**	**Gene 2**	**Gene 3**	**Gene 4**	**Gene 5**
One	ESTs Highly similar to hypothetical protein	EVL Enah/Vasp-like	Acetyl-Coenzyme A transporter	Extra spindle poles like 1	Major histocompatibility complex, class II, DO beta
Two	Cell division cycle 23, yeast, homolog	EVL Enah/Vasp-like	Extra spindle poles like 1 (S. cerevisiae)	Early growth response 1	Major histocompati-bility complex, class II, DO beta
Three	SRY-box 9 (sex determining region Y)-box 9	EVL Enah/Vasp-like	ESTs, Highly similar to hypothetical protein	C1q and tumor necrosis factor related protein 2	Homo sapiens mRNA from chromosome 5q21-22, clone:357Ex
Four	Cell division cycle 23, yeast, homolog	EVL Enah/Vasp-like	Protein associated with PRK1	Hypothetical protein MGC19556	Ubiquitin specific protease 9, Y chromosome (fat facets-like Drosophila)

Table 3 Prediction of chemo-response on initial biopsy samples

Tumor ID	Histologic response	Concordance with histological response
410	GR	No
197	PR	Yes
207	PR	Yes
278	GR	Yes
289	PR	Yes
345	GR	Yes
204	PR	Yes
274	PR	Yes
299	GR	Yes
464	PR	Yes
479	PR	Yes
481	PR	Yes
545	GR	Yes
654	GR	Yes