Identifying genes related to drug anticancer mechanisms using support vector machine

Lei Bao, Zhirong Sun*

Institute of Bioinformatics, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China

Received 28 February 2002; revised 6 May 2002; accepted 13 May 2002

First published online 24 May 2002

Edited by Robert B. Russell

Abstract In an effort to identify genes related to the cell line chemosensitivity and to evaluate the functional relationships between genes and anticancer drugs acting by the same mechanism, a supervised machine learning approach called support vector machine was used to label genes into any of the five predefined anticancer drug mechanistic categories. Among dozens of unequivocally categorized genes, many were known to be causally related to the drug mechanisms. For example, a few genes were found to be involved in the biological process triggered by the drugs (e.g. DNA polymerase epsilon was the direct target for the drugs from DNA antimetabolites category). DNA repair-related genes were found to be enriched for about eight-fold in the resulting gene set relative to the entire gene set. Some uncharacterized transcripts might be of interest in future studies. This method of correlating the drugs and genes provides a strategy for finding novel biologically significant relationships for molecular pharmacology. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Drug activity; Gene expression; Anticancer mechanism; Support vector machine

1. Introduction

The identification of drug–gene functional relationships is always an important issue in molecular pharmacology. Locating a drug target is the first step in rational drug design. Knowledge of the underlying genetic pathways affected by a drug enables better understanding of the drug's mechanism and the identification of gene markers for chemotherapy might be useful in clinical diagnosis. Recently, cDNA microarrays were used to assess gene expression profiles of 60 human cancer cell lines [1]. This gene expression database, together with the independently developed drug activity database recording anticancer profiles of various compounds against the same 60 cancer cell lines [2], provided the opportunity to take a global, systematic look at tumor molecular biology and pharmacology. The gene expression pattern or the drug activity pattern can be viewed as fingerprints for a gene or a drug that reflects their intrinsic properties. For example, drug activity pattern is known to be closely related to the drug's anticancer mechanism [3], and gene expression pattern is found to be closely related to the gene's biological function [4]. Whether a cell line is sensitive or resistant to a certain set of drugs is determined by what kind of genes it expresses and by how much these genes express. Therefore, the underlying causal drug–gene relationships might be visualized as the similarity between the two profiles. Based on this biological principle, Scherf et al. first raised the idea of associating drugs with genes using these two profiles [5]. An earlier work of Weinstein et al. denoted that drugs with similar anticancer mechanisms (within the same mechanistic category) have similar drug activity patterns, while drugs with different anticancer mechanisms (different mechanistic categories) have different drug activity patterns [3]. Therefore, each drug mechanistic category describes a unique probability distribution for the drug activity profile vectors. Mathematically, if we abstract both the gene expression profile and the drug activity profile as random vectors, we may treat them undiscriminatingly and give them a unified term as 'abstract profile vector'. The main assumption we hold is that, when the abstract profile vector of a gene is predicted with a high probability to be a sample drawn from one of the unique probability distributions (here, each probability distribution is equated with a certain mechanistic category), the gene and the drugs from that mechanistic category are probably biologically related. Based on this assumption, a newly developed machine learning method called support vector machine (SVM) [6] was used here to locate such drug–gene relationships. The drug activity profiles were used as training set to train the SVM and the gene expression profiles were used as test set to predict the mechanistic category they fell into. This paper analyzes whether drug activity patterns and gene expression patterns were able to identify genes related to the drug anticancer mechanism. Also, we aimed at providing for further experiments the uncharacterized gene candidates that might potentially be related to cancer chemotherapy efficacies.

2. Materials and methods

2.1. Materials

Gene expression was measured in 60 cell lines by microarrays and recorded as the base 2 logarithm of the ratios of the relative mRNA level of each cell line to a common reference sample pool [1]. Drug activity was expressed as the negative logarithm of GI_{50}, where GI_{50} was the concentration of the compound needed to cause 50% cell growth inhibition [7]. Weinstein et al. labeled 131 well-studied drugs into each of the six mechanistic categories: alkylating agents (Ak), topoisomerase I inhibitors (T1), topoisomerase II inhibitors (T2),

*Corresponding author. Fax: (86)-1062772237.
E-mail address: sunzhr@mail.tsinghua.edu.cn (Z. Sun).

Abbreviations: SVM, support vector machine; Ak, alkylating agents; T2, topoisomerase II inhibitors; Ri, RNA/DNA antimetabolites; Di, DNA antimetabolites; Mi, antimitotic agents

0014-5793/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(02)02835-1

RNA/DNA antimetabolites (Ri), DNA antimetabolites (Di) and antimitotic agents (Mi) [3]. The cell line set used for assessing the gene expression profiles and the cell line set used for assessing the drug activity profiles of the 131 drugs have 50 common members. Only data for these 50 cell lines was used for further calculation. 1400 compounds that have been well validated were selected according to Scherf et al. from a public database (http://discover.nci.nih.gov) [5]. The 1400-compound set was then combined with the 131-drug set and the union was further winnowed to eliminate those with more than five missing values. 1217 drugs passed this filter. 6357 genes with less than four missing values and with standard deviations larger than 0.5 were selected from the original database. The Unigene cluster identifier for each gene entry was extracted from the newest release of the SOURCE database (http://genome-www.stanford.edu/source). After removing chimera clones, repetitive clones and clones not included by the Unigene database, 4864 genes were reserved. Altogether, this work used a 1217×50 drug matrix and a 4864×50 gene matrix. Normalization was done so that every row of the two matrices has a 0 mean and a standard deviation of 1. The lists of final cell lines and drugs in mechanistic categories are available in Tables A1 and A2 of the Appendix.

2.2. Methods

For each mechanistic category, a corresponding binary SVM classifier was constructed. These classifiers were one-versus-rest classifiers ('one': positive class, 'rest': negative class). For each SVM classifier, compounds from each mechanistic category were used as the positive training set, while all the other compounds were used as the negative training set. Since the T1 category was comprised solely of camptothecin derivatives, this category was not used as a positive set. Therefore, altogether five binary SVM classifiers were built and trained (see Fig. 1 in Appendix).

A SVM requires the specification of two parameters: the kernel function and the penalty magnitude for violating the soft margin. The penalty magnitude was determined in the light of a trade-off between the sensitivity and the specificity. In this work, the drug mechanistic categories contained very few members relative to the total number of drugs in the data set. Therefore, there was an extreme imbalance between the number of positive and negative training examples. Without any modification, the SVM will misclassify all members of the training set as negative examples in the presence of noise. The problem was combated by modifying the diagonal of the kernel matrix during the training step according to Brown et al. [4]. For each positive example, a constant λ (n^+/N) was added to the diagonal entity while for each negative example, a constant λ (n^-/N) was added where N is the total number of training examples, n^+, n^- are the number of positive and negative training examples. The scale factor λ was set to 0.1. This method assigns a larger penalty to false negatives than to false positives. Polynomial kernel functions with powers of 1, 2, 3 and radial basis kernel were tested. For each drug mechanistic category, the training set was randomly split into 10 parts and a 10-fold cross-validation was carried out. To eliminate the effects introduced by a certain splitting, such splitting/validation procedures were repeated six times and the average cost for the six independent splittings was used to assess the optimal classification. The cost function to be optimized was defined as $fp+2fn$ where fp is the number of false positives and fn is the number of false negatives. Again, the false negatives were weighted more heavily because of the imbalance in thenumber of positive and negative training examples. After building the optimized SVM, a 4864-gene set was used as test set to predict their class labels.

To compare the frequency of certain function keywords in the predicted gene set and in the entire gene set, annotation record for each gene was also retrieved from the SOURCE database. Detail descriptions for the fields in each record can be found at http://genome-www5.stanford.edu/MicroArray/help/SOURCE/resultsBatchHelp.html. Altogether, four fields contained function annotation information. They are the fields of 'Summary Function', 'Gene Ontology Annotations', 'Other Ontology Annotations' and 'Enzyme Function'. We searched all these four fields for certain keywords and compared their frequency in the two gene sets. For the predicted gene set, each gene bearing the keywords was also artificially looked through to make sure that they are indeed involved in the expecting biological process to eliminate false positives. The keywords used for assessing genes involved in DNA repair were 'DNA' and 'REPAIR'; the keywords used for assessing genes involved in apoptosis and oncogenesis were 'APOPTOSIS' or 'APOPTOTIC' or 'ONCOGENE' or 'ONCOGENESIS'; the keywords used for assessing genes involved in protein synthesis were 'PROTEIN SYNTHESIS' or 'TRANSLATION' or 'RIBOSOMAL PROTEIN'. To eliminate the possibility that the enrichment for certain function keywords was simply due to more annotated genes in the predicted gene set than in the entire set, frequency normalized by the number of annotated genes was also calculated.

3. Results

3.1. Selection of the kernel function and the SVM accuracy

After filtering, the Ak, T2, Ri, Di and Mi categories (each positive training set) contained 34, 16, 17, 16 and 13 members. Each binary SVM classifier was constructed so as to represent the unique probability distribution for the corresponding mechanistic category. These classifiers were then used to predict which mechanistic category each gene would fall into according to its similarity to any of the five categories. If a gene did not fall into any of the five categories (all the discriminant values were less than zero), the gene was assigned to the 'unknown' category and the gene was not kept for further analysis. Given the data set, a proper kernel function and its parameters must be chosen to construct the SVM classifier. This selection is important because the type of kernel function determines the sample distribution in the mapping space. A first degree kernel is equal to a linear classifier. A second degree kernel reflects two-fold interactions between the measured data and so on for degree 3. There are no successful theoretical methods for determining the optimal kernel function and its parameters. After some preliminary computations, we found that as a whole the second degree polynomial kernel performed better than the first or third degree kernels or the radial kernel and is more appropriate in this context. Therefore, the second degree polynomial kernel was used to

Table 1
Average cost of six independent 10-fold cross-validation

Mechanistic categories	Different kernel functions			
	Polynomial 1	Polynomial 2	Polynomial 3	Radial basis
Ak	12.3	9.8	20.5	7.5
T2	9.8	6.2	13.5	6.7
Ri	20.7	13.7	17.5	14.3
Di	22.3	8.3	15.5	14.3
Mi	15.0	7.0	7.2	10.5
Total cost	80.1	45.0	74.2	53.3

The cost function is the number of false positive plus the double number of false negative.

Table 2
List of genes with positive labels

Alkylating agents (Ak)	MTHFD2, methylenetetrahydrofolate dehydrogenase
DCTN4, dynactin p62 subunit	SLC13A4, sulphate transporter 1
VAMP2, vesicle associated membrane protein 2	STK12, serine/threonine kinase 12
HBOA, histone acetyltransferase	GOT1, aspartate aminotransferase
FADD, mediator of receptor-induced toxicity	HEXA, hexosaminidase A
RAB5C, ras-related small GTP binding protein	TGFB1, transforming growth factor beta 1
DSP, desmoplakin	EST, SID 380833
SPINT2, placental bikunin	EST, SID 366842
EBI2, Chemokine (C-C) receptor 7	EST, SID 487257
LIG3, DNA ligase III	EST, SID 302936
PPFIBP1, liprin beta 1	EST, SID 469358
RECQL, DNA helicase Q1-like	EST, SID 153832
ANX11, Annexin XI	EST, SID 48294
EST, SID 486863	**DNA antimetabolites (Di)**
EST, SID 74554	RPA2, Replication protein A2
EST, SID 488184	LSP1, lymphocyte-specific protein
EST, SID 280229	PTDSS1, phosphatidylserine synthase I
EST, SID 31489	POLE, DNA polymerase epsilon
EST, SID 285992	RALGDS, guanine dissociation stimulator
EST, SID 381034	FABP5, fatty acid-binding protein
Topoisomerase II inhibitors (T2)	FTCD, formiminotransferase cyclodeaminase
ARRB2, beta-arrestin 2	CAPZA1, capping protein alpha
PTP4A2, protein-tyrosine phosphatase	SLC35A1, CMP-sialic acid transporter
NAP1L1, nucleosome assembly protein	GLO1, lactoylglutathione lyase
PDCD8, programmed cell death 8	MLH1, DNA mismatch repair protein
CBFB, core-binding factor	EST, SID 129023
MST4, serine/threonine-protein kinase	EST, SID 357231
GSTZ1, glutathione transferase Zeta 1	EST, SID 272950
EST, SID 429117	EST, SID 42114
EST, SID 287208	EST, SID 487981
EST, SID 62232	EST, SID 345634
EST, SID 133432	EST, SID 211325
EST, SID 127458	EST, SID 486727
EST, SID 143401	EST, SID 284962
RNA/DNA antimetabolites (Ri)	EST, SID 429908
EEF1B2, elongation factor 1	EST, SID 276816
HMG1, high-mobility group 1	**Antimitotic agents (Mi)**
RPL21, ribosomal protein L21	PDCD4, programmed cell death 4
RPL27, ribosomal protein L27	PFN1, profilin
AARS, Alanyl-tRNA synthetase	EST, SID 416406

Bold genes were the examples discussed in the text. For genes included by the GeneCard database, their GeneCard identifiers are given. For ESTs, the clone IDs were given.

Table 3
Different gene function distributions between the predicted set and the entire set

Function	Representative gene lists	Frequency in the predicted set (%)	Frequency in the entire set (%)	Enriched multiple
DNA Repair	LIG3 (Ak), RECQL (Ak), RPA2 (Di), POLE (Di), MLH1 (Di)	6.94 (12.20)	0.84 (1.62)	8.3 (7.5)
Apoptosis and oncogenesis	FADD (Ak), RAB5C (Ak), PDCD8 (T2), CBFB (T2), MLH1 (Di), RALGDS (Di), TGFB1 (Ri), PDCD4 (Mi)	10.67 (18.60)	4.85 (9.35)	2.2 (2.0)
Translation	EEF1B2 (Ri), RPL21 (Ri), RPL27 (Ri), AARS (Ri)	22.22 (36.36)	2.14 (4.12)	10.4 (8.8)

In column 2, predicted category is shown in the parentheses. In columns 3–5, numbers before the parentheses represent the frequency or multiple with respect to all the transcripts including ESTs, and numbers within the parentheses represent the frequency or multiple with respect to only the annotated genes (a subset of the former).

build the binary SVMs. The cross validation results are listed in Table 1.

3.2. Genes interpreting the common anticancer mechanism are enriched in the predicted gene set

All the genes having positive labels (namely, they do not belong to the 'unknown' category) are listed in Table 2. 19, 13, 18, 22 and 3 genes were assigned to the Ak, T2, Ri, Di and Mi categories respectively. Table 3 compares the frequency of some gene function definitions between the predicted gene set and the entire gene set (see the Appendix for more details). It is known that compounds from the Ak, T2, Ri and Di categories are able to induce DNA damage and genes involved in the DNA repair process are believed to affect the compound efficacies. Therefore, if our method was effective, DNA repair related genes were expected to be enriched in the predicted set than in the entire set. Table 3 shows that it is indeed the case. The DNA repair-related genes were enriched for about eight-fold in the predicted gene set, demonstrating the effectiveness of our approach to catch causal drug–gene relations. Another comparison was done for the genes involved in the apoptosis and oncogenesis process. The richness is not as significant as the DNA repair related genes, but we can still see a two-fold enrichment.

3.3. Individual causally related drug–gene examples

Several genes were closely related to the idiosyncracies of each mechanistic category. For example, the Ak category contains an apoptosis adaptor protein FADD. Overexpression of FADD sensitizes tumor cells to cisplatin (NSC 119875, and Ak agent) triggered cell death and cisplatin induced the recruitment of FADD to the apoptotic complex [8]. The Ri category included several genes involved in translation processes (Table 3). It is well known that the proportion between ribosomal proteins and rRNAs is precisely regulated; therefore, inhibiting RNA synthesis might also affect the ribosomal proteins as well as the protein synthesis process. As an example, a Ri agent 5-fluorouracil (NSC 19893) was known to inhibit the pre-rRNA processing and affect translation process [9]. Another example of the Ri category is the gene MTHFD2. MTHFD2 encodes a mitochondrial enzyme for tetrafolate metabolism and is involved in initiation of mitochondrial protein synthesis, a process affected by the antifolates from this category [10]. Very interestingly, the gene's cytosol counterpart MTHFD1 is known to be inhibited by the antifolates from this category [11]. An interesting gene from Di category is POLE, the DNA polymerase epsilon, which can be directly inhibited by some compounds from this category like ara-C (NSC 63878) [12]. Hence, POLE is an example of direct drug target found in this work. In the T2 category, the gene CBFB encodes a core-binding factor that plays multiple roles in the biological process such as apoptotic response and hematopoiesis. CBFB was known to be involved in chromosome inversion events and to form a fusion protein with gene MYH11 preferentially in topoisomerase inhibitor-treated patients [13]. In the Mi category, PFN1 encodes an actin binding protein which plays a role in the cell shaping [14]. It might be related to mitosis. Therefore, the correlation between PFN1 and the Mi was hypothesized.

4. Discussion

SVM has been successfully applied to categorize yeast gene according to their functions using gene expression profiles [4]. Here we extended that idea to identifying drug–gene relationships. The gene expression profile or the drug activity profile can be viewed as fingerprints for a gene or a drug. Weinstein et al. indicated that drug activity profiles were rich in information about mechanisms and each mechanistic category could be discriminated from the others [3]. The SVM cross validation results support this idea. For four out of five categories, the defined costs for the optimized classifiers were less than 10. SVM has several advantages over such machine learning problems. First, SVM avoids overfitting and finding trivial results (a common problem in machine learning field) by implementing the structural risk minimization principle. Second, SVM condenses the information of the training samples into a small number of samples called support vectors. If all the other training samples are removed and SVM is retrained, the solution would keep unchanged. This allows SVM to classify new examples efficiently, since the majority of the training examples can be safely ignored. Third, the solution of SVM was not affected by the initial conditions. Therefore, it is quite easy to be realized.

In this paper, we have demonstrated that SVM can be used to determine drug–gene functional relationships with both relatively high sensitivity and high specificity. Since the gene expression profiles are those for untreated cells, the relationships established between drugs and genes should first be considered correlative and not definitely causal. However, the literature indicated that the drug–gene groups found here were rich in causal relationships. It is not surprising, both biologically and algorithmically. Biologically, without regard to experiment errors, it seems intuitively reasonable to presume that a gene is probably related to a drug mechanistic category if the gene's expression pattern is similar to the drug activity patterns for that category. The intrinsic factors for the cell lines behavior to compounds acting by the same mecha-

nism are the genes involved in the biological process underlying the drugs' anticancer efficacies. Furthermore, anecdotal evidence exists that causally related drug–gene pairs exhibit similar profiles [5]. Algorithmically, in many cases SVM outperforms other machine learning methods like neural networks. Scherf et al. first used average-linkage clustering to associate genes and drugs, but they had difficulty locating true functional correlations from the noise [5]. Butte et al. used the permutation method to deduce a very stringent correlation coefficient threshold [15]. Although they did find one hypothetical gene–drug pair that passed the threshold, the sensitivity seemed too low (one hypothetical functional pair out of 33 million drug–gene pairs). The similarity measure here was more than just linear correlations. Most importantly, as opposed to these two approaches that associate a single gene with a single drug during the first step, our approach associated a group of genes with a group of drugs acting by the same mechanism. Bringing genes into the context of drug mechanism will give more information on the gene functions than simply correlating drugs and genes. For unknown ESTs, this approach provides useful clues for their roles in cancer occurrence and chemotherapy; therefore, in some sense, this approach may also be used in a limited way to predict gene functions. Certain ESTs are indicated as deserving high priority in future molecular studies. Together with other information like sequence features, subcellular locations and so on, the functions of these ESTs might be disclosed.

There are also several limitations for this method. First, since only the gene expression profiles were taken into account, our method can work only when the drug–gene functional relationships were embodied at the transcriptional level. Given the fact that drug–gene functional relationships may take place at various levels, this method can only identify a subset of all the potential causally relationships. Second, this approach is a supervised learning method. The major limitation is that a priori knowledge of drug mechanisms is required. These five categories were well studied mechanistic categories. Although they account for only a small proportion of the anticancer compound database, yet with more and more mechanistic categories identified, this approach would exert better use.

Acknowledgements: The authors would like to thank Professor Patrick O. Brown et al. and John Weinstein et al. for proving the GI50 data and gene expression data and thank Professor David Haussler for his SVM software and helpful suggestions. This work is supported by the National Natural Science Grant (China) (No. 39980007 and No. 19947006).

Appendix

Table A1
50 cancer cell lines used in this work

Tissue	Cell lines
CNS	SNB-19, SNB-75 SF-268, SF-295, SF-539, U251
CO	COLO205, HCC-2998, HCT-116, HCT-15, HT29, KM12, SW-620
LC	A549/ATCC, EKVX, HOP-62, HOP-92, NCI-H226, NCI-H23, NCI-H322M, NCI-H460, NCI-H522
LE	CCRF-CEM, HL-60, K-562, MOLT-4, RPMI-8226, SR
ME	LOXIMVI, M14, MALME-3M, SK-MEL-2, SK-MEL-5, SK-MEL-28, UACC-62, UACC-257
OV	IGROV1, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, SK-OV-3
RE	786-0, A498, ACHN, CAKI-1 RXF-393, SN12C, TK-10, UO-31

The tissue origin of 50 cell lines was expressed here in abbreviations. CNS: Central nervous system; CO: colon; LC: lung; LE: leukemia; ME: melanoma; OV: ovarian; RE: renal.

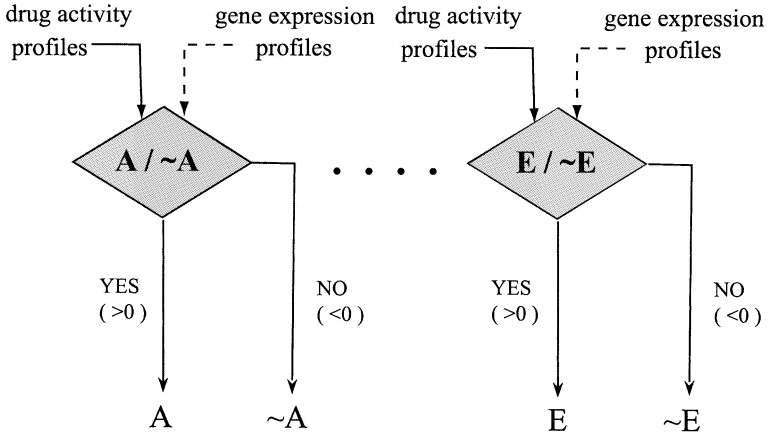

Fig. 1. Illustration of the method. A–E: Denotes the five drug mechanistic categories: Ak, T2, Ri, Di, Mi. In the training phase, inputs (solid arrows) are the drug activity profiles (drug matrix). Take A/~A classifier as an example, drugs within category A serve as positive training samples, while all the other drugs serve as negative training samples (~A category indicates they do not belong to the category A). Altogether five parallel SVM classifiers were trained (A–E). In the predicting phase, inputs (dashed arrows) are the gene expression profiles (gene matrix). For each input gene, each of the five trained SVM classifiers will give a scalar output – the discriminant value. If the maximum of the five outputs is larger than 0, the gene is assigned to the corresponding drug mechanistic category; otherwise if all the five outputs are smaller than 0, the gene is assigned to the 'unknown' category (a sixth category which is uninteresting). Most genes were found to fall into the 'unknown' category.

Table A2
Five mechanistic categories serving as five positive training sets

Mechanistic categories	Anticancer drug NSC number
Ak (34)	750, 762, 3088, 6396, 8806, 9706, 25154, 26980, 34462, 56410, 73754, 79037, 95441, 95466, 102627, 119875, 132313, 135758, 142982, 167780, 172112, 182986, 241240, 256927, 271674, 296934, 329680, 338947, 344007, 348948, 353451, 357704, 363812, 409962
T2 (16)	82151, 122819, 123127, 141540, 164011, 249992, 267469, 268242, 269148, 301739, 308847, 337766, 349174, 354646, 355644, 366140
Ri (17)	740, 19893, 102816, 126771, 132483, 139105, 143095, 148958, 153353, 163501, 174121, 184692, 224131, 264880, 352122, 368390, 633713
Di (16)	752, 755, 1895, 27640, 32065, 51143, 63878, 71261, 71851, 95678, 107392, 118994, 127716, 145668, 303812, 330500
Mi (13)	757, 33410, 49842, 67574, 83265, 125973, 153858, 332598, 361792, 376128, 406042, 608832, 609395

Table A3
Frequency of the gene function keywords in the predicted set and entire set

Function	Related categories	Predicted gene set			Entire gene set		
		positive[a]	known[b]	all[c]	positive[a]	known[b]	all[c]
DNA repair	Ak, T2, Ri, Di	5	41	72	41	2524	4864
Apoptosis/oncogenesis	Ak, T2, Ri, Di, Mi	8	43	75	236	2524	4864
Protein synthesis	Ri	4	11	18	104	2524	4864

[a]Number of genes with function specified in the first column.
[b]Number of genes having function annotations.
[c]Total number of the genes including ESTs.

References

[1] Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., VandeRijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. (2000) Nat. Genet. 24, 227–235.
[2] Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace Jr., A.J., Kohn,, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E. and Paull, K.D. (1997) Science 275, 343–349.
[3] Weinstein, J.N., Kohn, K.W., Grever, M.R., Viswanadhan, V.N., Rubinstein, L.V., Monks, A.P., Scudiero, D.A., Welch, L., Koutsoukos, A.D., Chiausa, A.J. and Paull, K.D. (1992) Science 258, 447–451.
[4] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares Jr., M. and Haussler, D. (2000) Proc. Natl. Acad. Sci. USA 97, 262–267.
[5] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O. and Weinstein, J.N. (2000) Nat. Genet. 24, 236–244.
[6] Cortes, C. and Vapnik, V. (1995) Mach. Learn. 20, 273–293.
[7] Boyd, M.R. and Paull, K.D. (1995) Drug Dev. Res. 34, 91–109.
[8] Micheau, O., Solary, E., Hammann, A. and Dimanche-Boitrel, M.T. (1999) J. Biol. Chem. 274, 7987–7992.
[9] Ghoshal, K. and Jacob, S.T. (1997) Biochem. Pharmacol. 53, 1569–1575.
[10] Wallace, R.B. and Freeman, K.B. (1974) Biochim. Biophys. Acta 366, 466–473.
[11] Hollinger, J.L., Hommes, O.R., van de Wiel, T.J., Kok, J.C. and Jansen, M.J. (1982) J. Neurochem. 38, 638–642.
[12] Mirzayans, R., Dietrich, K. and Paterson, M.C. (1993) Carcinogenesis 14, 2621–2626.
[13] Dissing, M., Le Beau, M.M. and Pedersen-Bjergaard, J. (1998) J. Clin. Oncol. 16, 1890–1896.
[14] Witke, W., Sutherland, J.D., Sharpe, A., Arai, M. and Kwiatkowski, D.J. (2001) Proc. Natl. Acad. Sci. USA 98, 3832–3836.
[15] Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. and Kohane, I.S. (2000) Proc. Natl. Acad. Sci. USA 97, 12182–12186.