ELSEVIER

Biochemical Pharmacology 62 (2001) 547–553

Biochemical Pharmacology

Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression database☆

Giuseppe Musumarra*, Daniele F. Condorelli, Salvatore Scire, Alessandro S. Costa

Dipartimento di Scienze Chimiche, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy

Received 8 January 2001; accepted 26 April 2001

Abstract

Application of a soft multivariate statistical procedure, called PLS, partial least squares modelling in latent variables or projections to latent structures, allows extensive exploitation of the enormous amount of information embedded in the National Cancer Institute gene expression and antitumour screen databases. Interpretation of the statistical results provides new significant biological insights such as classification of human tumour cell lines based on their gene expression patterns, evaluation of the influence of gene transcripts on drug efficacy and assessment of their selectivity for classes of compounds which act by the same mechanism, and identification of uncharacterized gene expression targets involved in cancer chemotherapy. Among them, the transcripts GC11121, GC17689, and GC18564 (unknown gene products extremely selective for RNA/DNA antimetabolites) are indicated by the present work as deserving high priority in future molecular studies. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Antitumour agents; Multivariate analysis; Partial least squares; Biometrics; Gene expression targets; NCI database

1. Introduction

Future hope in the effectiveness of cancer chemotherapy is based on the general acceptance that genetic events underlie the development of cancer. In spite of paramount advances in genome-scale research, the goal of a molecular-based cancer therapy is far from being achieved. In this context, a prodigious amount of information accumulated by the National Cancer Institute has recently been provided [1] to the scientific community as a gene expression database including 9703 cDNAs or gene transcripts representing ~8,000 unique genes among 60 human tumour cell lines used in a drug discovery *in vitro* screening, including drug activity profiles for leukaemia, non-small-cell lung, central nervous system, colon, melanoma, ovarian, and renal tumour cell lines. The novel technology of cDNA microarrays [2] allowed the rapid generation of a large amount of data, representing an estimate of the levels of specific mRNAs transcribed from thousands of different genes. The above database represents a starting point for understanding the complex relationships between gene expression and drug activity. Achievement of the above ambitious task requires investment of unprecedented efforts and resources in studies that have necessarily to be guided by strategies identifying shortcuts. The key role of bioinformatics in answering a few preliminary questions before undertaking labour intensive studies is now being recognised, but not yet fully exploited. A few important questions to be answered are: what are the main gene expression targets involved in cancer chemotherapy for classes of drugs acting by known mechanisms (e.g. topoisomerase II inhibitors, etc.) or for newly discovered drugs? Which are the most important gene expression targets involved in cancer chemotherapy not yet fully characterised that need to be assigned priority in further studies?

Scherf *et al*. [3] recently reported an application linking bioinformatics and chemoinformatics by correlating gene

☆**Supplementary information** is available on *Elsevier's* World Wide Web site (http://www.elsevier.nl) or from the corresponding authors.

* Corresponding author. Tel.: +39-095-334-175; fax: +39-095-580-138.

E-mail address: gmusumarra@dipchi.unict.it (G. Musumarra).

Abbreviations: NCI, National Cancer Institute; PLS, partial least squares modelling in latent variables or projections to latent structures; SIMCA, soft independent modelling of class analogy; PCA, principal component analysis; PC, principal components; and VIP, variable importance in the projection.

PII: S0006-2952(01)00711-0

expression and drug activity patterns in the NCI 60 cell lines with the aim "to provide a rationale for selection of therapy on the basis of molecular characteristics of a patient's tumour" using a statistical approach based on cluster analysis and Pearson correlation coefficients. Cell–cell correlations based on gene expression profiles (T matrix) and on drug activity profiles (A matrix) provided gene–drug correlations (A-T matrix clustering) [3].

An alternative and more powerful multivariate statistical procedure, called PLS and included in the SIMCA package [4], is aimed at finding relationships between a group of explanatory variables (the X matrix including the "descriptors") and a set of dependent variables (the Y matrix including the "responses"). In the past two decades, the SIMCA method [4–7] has been successfully applied in many fields of science and has been demonstrated to be a very powerful approach to handle complex data sets represented in the form of matrices where a number of objects are characterised by a number of variables. The main advantage of PLS is that it investigates the relationships among all objects and all variables simultaneously by means of PCA in both X and Y matrices under the constraint that the PC extracted from each matrix are linearly correlated to each other. Application of the above "soft" statistical methodology allows to relate the expression level of thousands of different genes (called the "descriptors" in the PLS procedure) to the therapeutical "fingerprints" of a set of compounds (called the "responses" in the PLS procedure) for the same cell lines (objects), evaluating the influence of each gene expression target in determining the therapeutical responses. We have recently reported [8] the first application of the PLS methodology to compounds whose molecular targets are well known and pointed out the consistency of the statistical results with experimental evidence reported in the literature.

The purpose of the present work was to provide possible shortcuts by suggesting priorities in future genome pharmacology studies aimed at the identification of the main uncharacterised gene expression targets involved in cancer chemotherapy. Discussion of the functional relationships between known gene products and drug activities, deserving specific detailed comparisons, will be limited to a few examples just to show the consistency of the statistical results with current knowledge in the field.

2. Materials and methods

The data set used for PCA was a table (matrix) in which 60 cell lines were characterized by multivariate biological "fingerprints" given by the gene expression profiles. The results of PCA depend upon the weighting of the data; in the present case the variables were autoscaled by multiplying the variables by appropriate weights (the reciprocal of the variable standard deviation) to give them unit variance (i.e. the same importance).

PCA was carried out by means of SIMCA software package [4] on a data matrix containing 576300 (9605 × 60) elements X_{ik}, where index k is used for the gene expression profiles (variables) and index i for the cell lines (objects). Autoscaled matrix elements were then fitted into a model given by equation (1), where the number A of significant cross-terms (components), and the parameters p_{ak} and t_{ia} are calculated by minimising the residuals e_{ik}, after subtracting $\bar{x}_k$ (the mean value of the i^{th} experimental quantities x_{ik}).

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{a=A} t_{ia}p_{ak} + e_{ik} \tag{1}$$

Parameters $\bar{x}_k$ and p_{ak} (the loadings) depend only on the gene expression profiles (variables), and the t_{ia} (scores) only on the cell lines.

The deviations from the model are expressed by the residuals e_{ik}. The number of significant components (A) was determined using the cross-validation technique [7].

Relationships between two blocks of variables, the "descriptor" matrix X and the "response" matrix Y, can be achieved by PLS analysis [9–11], where the members of the Y matrix can be described as a function of the members of the X matrix. The PLS algorithm computes PLS components for each of the two matrices looking simultaneously for a linear relationship between the X-scores (t_{ia}) and the Y-scores (u_{ia}) reported in equation (2), which is the analogous to equation (1) for the Y matrix:

$$y_{im} = \bar{y}_m + \sum_{a=1}^{a=A} u_{ia}c_{am} + g_{im} \tag{2}$$

The algorithm is iterative for each dimension as in PCA and consists in finding the latent variables of the X and Y matrices in such a way that the relationship between t_{ia} and u_{ia} is maximised.

The statistical results obtained by the PLS method are able to detect what variables in the X block are relevant to determine the dependent variables (Y block) by means of the VIP values. The VIP values reflect, in fact, the importance of terms in the model both with respect to Y, i.e. its correlation to all the responses, and with respect to X. SIMCA computes VIP values [4] by summing over all model dimensions the contributions VIN (variable influence). For a given PLS dimension, a, $(\mathrm{VIN})_{ak}$ [2] is equal to the squared PLS weight $(w_{ak})^2$ of that term, multiplied by the percent explained of residual sum of squares by that PLS dimension. The accumulated (over all PLS dimensions) value, $VIP_k = \Sigma_a\, (VIN)_k^2$ is then divided by the total percent explained of residual sum of squares by the PLS model and multiplied by the number of terms in the model.

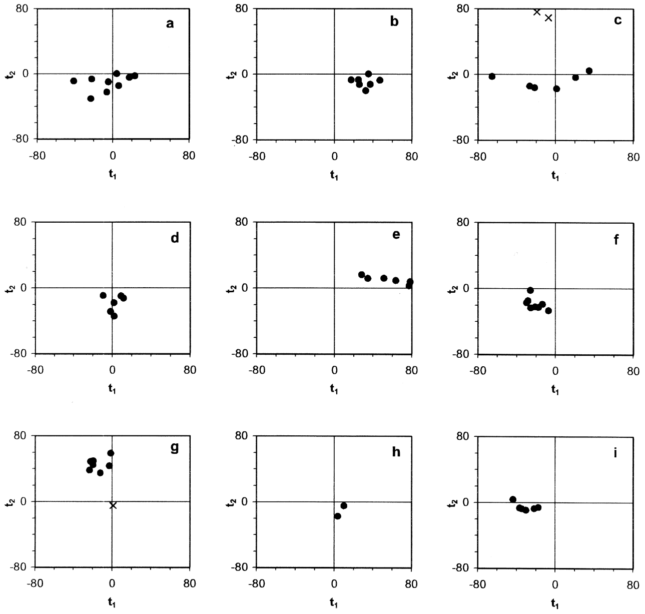

Fig. 1. t_1–t_2 scores plot from PCA using 60 cell lines as objects and 9605 gene expression profiles as variables (two PC explain already 15.7% variance): a, non-small-cell lung; b, colon; c, breast; d, ovarian; e, leukaemia; f, renal; g, melanoma; h: prostate; i, central nervous system. Outlier cell lines excluded in the PLS analysis (see text) are designated as x in plots c and g.

3. Results

The NCI *in vitro* antitumour screen database can be represented in the form of a matrix, where the 60 cell lines (i.e. objects) may be represented as characterised by a multivariate biological fingerprint, given by the gene expression profiles (the "descriptor" variables in the X matrix) or by multivariate therapeutical fingerprints given by drug activity patterns expressed as log GI_{50} [1] (the "response" variables in the Y matrix). Both X and Y matrices can undergo separate PCA based on gene transcript levels (X matrix) or on drug antitumour activities (Y matrix). No cell line clustering according to histological origin based on drug responses has been previously noted either by cluster analysis [3] or PCA [8]. On the contrary, cell line grouping by organ of origin has been evidenced by cluster analysis of gene expression data [3,12]. PCA carried out in the present work on 9605 gene expression data (X matrix, where 98 genes were excluded from the original 9703 gene transcripts in the above cited database due to a high number of missing data) gave a 4PC significant model explaining 25.2% of variance, providing cell line classification (see t_1–t_2 scores plot in Fig. 1) consistent with that previously reported [3,12]. In particular, two breast cell lines (MDA-MB435 and MDA-N) and one melanoma (LOX IMVI) are clearly outliers in the t_1–t_2 scores plot with respect to other cancer cells with the same origin (Fig. 1, c and g). This result supports the hypothesis of Ross *et al*. [12] that MDA-MB435 and MDA-N cell lines derived from a single patient with breast cancer possibly originated from a co-existing occult melanoma. Taking into account the results of PCA commented on above, three cell lines (MDA-MB435, MDA-N, and LOX IMVI) were excluded from further analyses. Therefore, all PLS models reported below include only 57 cell lines (objects) characterised by gene expression profiles in X matrices (descriptors) or by drug antitumour activities expressed as log GI_{50} in Y matrices (responses). The X matrices included 9605 variables in all PLS models, while the Y matrices included a number of variables ranging from 171 for the model with all drugs to 6 for the model with antimitotic drugs only (for details see Table 1). PLS correlates the block of descriptors to the block of therapeutical responses for the same cell lines by simultaneously considering all descriptor and response variables, providing models characterised by statistical parameters such as the number of PLS components, the percentage of variance explained in both X and Y matrices, the predicting ability (Q^2), as well as the influence of each gene transcript in determining the therapeutical responses

Table 1
Statistical parameters for PLS models 1–6[a]

Entry	Model	No. of variables	No. of objects	No. of PLS components	Q^2 (%)	Variance explained (%)
1	All	X 9605 Y 171	57	4	17.2	22.1 (8.5 + 5.0 + 4.6 + 4.0) 39.5 (18.7 + 11.2 + 4.8 + 4.8)
2	T	X 9605 Y 16	57	2	25.3	12.9 (6.8 + 6.1) 54.6 (37.8 + 16.8)
3	R	X 9605 Y 13	57	2	25.6	13.1 (8.6 + 4.5) 48.5 (32.4 + 16.1)
4	D	X 9605 Y 16	57	2	22.8	13.0 (8.8 + 4.2) 46.8 (25.2 + 21.6)
5	A	X 9605 Y 36	57	3	31.1	17.2 (7.1 + 6.3 + 3.8) 62.9 (38.9 + 15.9 + 8.1)
6	M	X 9605 Y 6	57	3	18.3	16.5 (7.2 + 5.5 + 3.8) 67.9 (34.9 + 20.9 + 12.1)

[a] Drug class designation: T = topoisomerase II inhibitors, R = RNA/DNA antimetabolites, D = DNA antimetabolites, A = alkylating agents, M = antimitotic agents. The full list of drugs is available as supplementary information.

(VIP). The latter value represents a statistical parameter ranking all descriptors in order of decreasing importance.

Different empirical PLS models were derived by selecting an appropriate number of cell lines (objects) and therapeutical responses (variables in the Y matrix), while the number of descriptor variables (9605) was the same throughout the analysis. The results of all PLS models are summarised in Table 1.

PLS parameters for model 1 provide an overall insight into the capability of the 9605 molecular target genes to account for the therapeutical ability of 171 drugs with known and unknown mechanism taken from the *in vitro* anticancer screening standard database [1]. 4 PLS components explain 39.5% variance of the therapeutical response matrix for the considered 57 cell lines, confirming that the considered gene expression profiles represent a predominant factor in determining the drug activities. However, in order to answer the first question (what are the main gene expression targets involved in cancer chemotherapy?), the order of importance of each descriptor on biological responses has to be evaluated for more homogeneous drug subsets. Each of models 2–6, derived by relating the 9605 descriptors to a class of drugs acting by the same mechanism (T = topoisomerase II inhibitors, R = RNA/DNA antimetabolites, D = DNA antimetabolites, A = alkylating agents, M = antimitotic agents respectively), exhibits very interesting statistical parameters, recorded in Table 1. Only 2 or 3 PLS components are required to obtain statistically significant PLS models explaining percentages of Y matrix variances ranging from 46.8% up to 67.9%. This finding confirms on a sound statistical basis the relevance of gene expressions in determining the activities for each class of drugs. The VIP values for models 2–6, all using the same X matrix with the same 9605 descriptors, represent a proper statistical parameter to select the main gene expression targets involved in cancer chemotherapy for a specific class of drugs (i.e. topoisomerase II inhibitors). Fig. 2 provides, as an example, a graphical picture of the VIPs for model 2 (T = topoisomerase II inhibitors), while a complete list of all 9605 genes in order of importance for each PLS model is available in the www.elsevier.nl as supplementary information. A proper statistical criterion would suggest to discuss all descriptor variables exhibiting VIPs above a given value. However, adopting an alternative arbitrary option dictated by the need of simplicity and conciseness and suggested by a most popular term used to identify “hit parades” (CD, movies, books, rich men, etc.), we here report (Table 2) the “top ten” gene transcripts (designated by the NCI database identification number) selected by each PLS model from the original set of 9605.

The second column of Table 3, indicating the common names for the same gene products reported in Table 2, points out that 13 out of the 44 “top ten” genes encode for an unknown product. This finding provides an answer to the second question: which are the most important gene expression targets involved in cancer chemotherapy not yet fully characterised that need to be assigned priority in further studies?

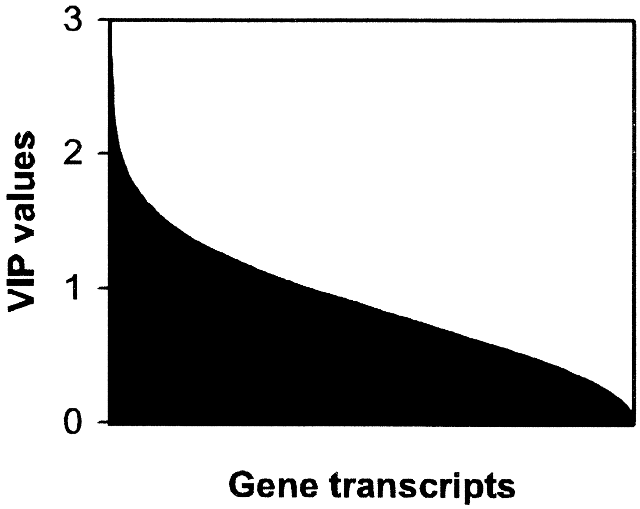

Fig. 2. VIP plot for model 2 (topoisomerase II inhibitors); 149 gene transcripts exhibit VIP values > 2.

4. Discussion

In addition to the identification of clear guidelines in establishing priorities for further gene structural and functional investigations, the results of the PLS analysis can be related to other biologically relevant matters. In order to compare the influence of each gene in models 2–6, the rank order of all top ten genes in the VIP lists for models 2–6 is also reported in columns 3–7 of Table 3. An analysis of these results allows identification of gene transcripts that are selectively influencing the activity of a single class of drugs, as defined by the mechanism of action. However, several genes are able to simultaneously exert a powerful influence on the activities of multiple classes of drugs. Of course, these genes should be involved in cellular and molecular processes that are common to the mechanism of action or of resistance of drugs belonging to different classes.

A distinction between M drugs and other classes of compounds (T, R, D, A) is evident by inspecting Table 3. Genes exerting an important influence on M drug action (9 out of the top 10 in the VIP list, with exception of GC17402) are located in a low rank order position in the VIP lists referring to other drug classes: such genes are designed hereinafter as M+. On the contrary, it is possible to distinguish gene products (GC10461, GC11800, GC13601, GC14119, GC16251, GC16502) influencing the action of all classes of drugs with the exception of M drugs. It is not surprising that several genes simultaneously influence the activity of T, R, D, A drugs: all these compounds induce DNA damage through different mechanisms and it can be predicted that several proteins involved in DNA replication and repair are able to affect their efficacy. In

Table 2
Top ten VIP lists for gene expression profiles for models 2–6[a]

Rank	T	R	D	A	M
1st	GC14400	GC15528	GC10461	GC10461	GC11379
2nd	GC17167	GC10461	GC11800	GC11800	GC13124
3rd	GC14538	GC18781	GC16502	GC18530	GC14113
4th	GC16345	GC17402	GC19391	GC16468	GC18981
5th	GC16468	GC11121	GC12935	GC10079	GC11486
6th	GC12274	GC15605	GC13261	GC13617	GC10442
7th	GC17315	GC18564	GC13399	GC11319	GC17402
8th	GC11319	GC14119	GC16251	GC15198	GC10867
9th	GC10903	GC17689	GC17472	GC10728	GC10246
10th	GC18174	GC13601	GC15354	GC11492	GC15718

[a] Drug class designation as in Table 1.

spite of the similarities in the mechanism of action of T, R, D, A drugs, it is still possible to clearly identify gene transcripts that show a striking level of selectivity for a single class. For instance, the transcripts GC11121, GC17689, and GC18564 (unknown gene products extremely selective for R class) should in our opinion deserve high priority for further studies.

An interesting example of M+ gene is the p21-activated kinase 2 (PAK2), a ubiquitously expressed serine/threonine kinase (Table 3). PAK2 is member of a family of highly conserved kinases that are activated by interactions with Rac1 and Cdc42 [13]. Current evidence suggests that PAKs are involved in regulating some of the diverse actin cytoskeleton changes induced by Rac and Cdc42 [14]. Indeed, controlled changes to the actin cytoskeleton are vital for almost all cellular processes including motility, adhesion, cell division, and cell death. It has been reported that Rac1, one of the activators of PAKs, is required for cell proliferation and G2/M progression [15], suggesting that the cytoskeletal alterations that need to occur before mitosis and cytokinesis require the participation of Rac proteins. If PAK2 is mediating the effects of Rac1 on G2/M progression, it would be clear why the level of expression of this kinase is not influencing the action of drugs (T, A, D, R) that induce a cell cycle arrest in G0/G1. On the other hand, the action of antimitotic (M) drugs that act on the mitotic spindle in the M phase, such as vinblastine, can be influenced by molecular events that take place at the G2/M phase. In accordance with an influence of the actin cytoskeleton on M drug action, another M+ gene encodes for profilin-1, a protein that binds to actin and affects the structure of the cytoskeleton. Although several other alternative mechanisms, apart from interactions with the actin cytoskeleton, could explain the influence of PAK2 and profilin on sensitivity to M drugs, the results of the present multivariate statistical analysis suggest carrying out further experimental work to establish a functional link between these proteins and M drug action.

The gene transcript with the highest VIP value for R drugs (Table 3) encodes for the NAD-dependent methylenetetrahydrofolate dehydrogenase (MTHFD2), a nuclear-encoded mitochondrial bifunctional enzyme with methylenetetrahydrofolate dehydrogenase and methenyltetrahydrofolate cyclohydrolase activities [16]. Its role is to provide formyltetrahydrofolate for the synthesis of formylmethionyl transfer RNA required for the initiation of protein synthesis in mitochondria. It is expressed in transformed or established mammalian cell lines *in vitro* but not in most adult tissues [17]. The gene in quiescent Balb/c 3T3 fibroblasts is induced by mitogens such as serum and phorbol esters and requires *de novo* transcription. The intracellular location of the enzyme and its regulation of expression are consistent with its proposed role in mitochondrial biogenesis. The correlation between the transcription of this enzyme and the susceptibility to the effects of R drugs might suggest that mitochondrial transcription may represent an important molecular target for this class of antitumor agents. Alternatively, an increased mitochondrial biogenesis might be correlated with a global increase in cellular RNA synthesis that would account for the higher sensitivity to the action of R drugs. Interestingly, another gene transcript influencing R drug efficacy (GC18781) encodes for a subunit of a pre-mRNA splicing factor SP2, localized predominantly in the mitochondrial matrix and putatively involved in nucleus–mitochondrion interactions.

In conclusion, application of multivariate statistical procedures such as PCA and PLS allows extensive exploitation of the enormous amount of information embedded in the NCI gene expression and antitumour screen databases. Interpretation of the statistical results throws light on biologically relevant problems such as: (a) classification of human cell lines according to the tissue of origin based on their gene expression patterns; (b) evaluation of the influence of gene transcripts measured by cDNA microarrays (including both well known ones and not fully characterized gene products) on the sensitivity to drug treatment for classes of compounds which act by the same mechanism; (c) indication of possible shortcuts for future molecular studies aimed at the identification of the main uncharacterized gene expression targets involved in cancer chemotherapy.

Table 3
Ranks of gene transcripts reported in Table 2 in VIP lists for models 2–6[a]

ID number	Gene product[b]	T	R	D	A	M
GC10079	unknown	933	6590	25	5	3273
GC10246	PFN1	6249	1364	6571	6704	9
GC10442	KIAA0896	5160	4783	3323	5237	6
GC10461	TUBB	29	2	1	1	5863
GC10728	KIAA0165	18	1405	38	9	294
GC10867	RPS18	8006	3529	4542	8433	8
GC10903	KIAA0957	9	1590	1573	519	4401
GC11121	unknown	2535	5	905	3582	7305
GC11319	EVI5 (NB4S)	8	439	35	7	1650
GC11379	SRD5A1	1175	7028	5909	5122	1
GC11486	unknown	3053	1097	4004	1908	5
GC11492	SLA	79	611	62	10	1566
GC11800	unknown	19	298	2	2	2815
GC12274	STX3A	6	5552	897	65	1627
GC12935	unknown	1006	48	5	292	2680
GC13124	ABCC6 (MRP6)	1043	3317	4788	1527	2
GC13261	TGFBR3	37	4109	6	22	8841
GC13399	unknown	774	21	7	349	1675
GC13601	PLK	834	10	24	787	2930
GC13617	WAS	198	652	30	6	1146
GC14113	PAK2	6227	1472	2958	8249	3
GC14119	WHSC2	303	8	15	327	4820
GC14400	DKFZP586F1918	1	5841	513	34	1207
GC14538	C17orflB	3	3100	167	45	5606
GC15198	ZNF184	15	379	50	8	1212
GC15354	TMSB4X	320	1022	10	286	1567
GC15528	MTHFD2	249	1	1742	1351	130
GC15605	unknown	338	6	960	571	1933
GC15718	unknown	3563	4286	892	3653	10
GC16251	HYA22	246	478	8	17	2410
GC16345	MPDU1	4	82	26	19	1711
GC16468	PIX	5	2927	58	4	863
GC16502	HYA22	49	1299	3	89	3006
GC17167	PPP4C	2	909	2101	293	9065
GC17315	CD3D	7	888	169	24	747
GC17402	REA	1651	4	472	1464	7
GC17472	unknown	44	31	9	50	753
GC17689	unknown	5646	9	3600	2599	4991
GC18174	TUFM	10	65	895	74	63
GC18530	FAK	12	1688	343	3	858
GC18564	Unknown	873	7	647	1183	1405
GC18781	C1QBP	400	3	914	707	527
GC18981	Unknown	394	3903	2833	7214	4
GC19391	Unknown	6724	2090	4	717	8453

[a] drug class designation as in Table 1.
[b] nomenclature according to Ref. 18.

The present results point out the key role of the PLS multivariate approach in the identification of strategies for future genome-scale cancer pharmacology studies.

Acknowledgments

We wish to thank Dr. Vincenzo Costa for writing a program which converted the data base into a SIMCA data file. Financial support of the University of Catania, of MURST, and CNR (Roma) are gratefully acknowledged.

References

[1] Database resources currently available on the World Wide Web: http://www.dtp.nci.nih.gov/.

[2] Duggan DJ, Bittner Y, Chen P, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999;21:10–4.

[3] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN, Scherf U. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–44.

[4] SIMCA-P 8.0, User Guide and Tutorial, Umetri AB, Umeå, Sweden, 1999.

[5] Wold S, Sjöström M. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. American Chemical Society Symposium Series 1977;52:243–82.

[6] Wold S. Principal component analysis. Chemometr Intell Lab Syst 1987;2:37–52.

[7] Wold S. Cross-validatory estimation of the number of components in factor and principal component models. Technometrics 1978;20:397–405.

[8] Musumarra G, Condorelli DF, Costa AS, Fichera M. A multivariate insight into the *in vitro* antitumor screen database of the National Cancer Institute: classification of compounds, similarities among cell lines and influence of molecular targets. J Comput Aid Mol Des 2001;15:219–34.

[9] Wold S, Albano C, Dunn WJ, Esbensen K, Hellberg S, Johannsson E, Sjöström M. Pattern recognition: finding and using regularities in multivariate data. Martens H, Russwurm H, editors. Food Research and Data Analysis. London: Applied Science, 1983. p. 147–67.

[10] Wold S, Albano C, Dunn WJ, Edlund U, Esbensen K, Geladi P, Hellberg S, Lindenberg W, Sjöström M. Multivariate data analysis in chemistry, In: Kowalski BR, editor. Chemical Mathematics and Statistics in Chemistry, NATO ASI Series C. Dordrecht: D. Reidel Publ. Co., 1984. p. 17–95.

[11] Clementi S, Cruciani G, Curti G. Some applications of the Partial Least Squares method. Anal Chimica Acta 1986;191:149–60.

[12] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van De Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227–35.

[13] Bagrodia S, Cerione R. PAK to the Future. Trends in Cell Biology 1999;9:350–5.

[14] Daniels RH, Bokoch GM. p21-activated protein kinase: a crucial component of morphological signaling?. Trends Biochem Sci 1999; 24:350–5.

[15] Moore KA, Sethi R, Doanes AM, Johnson TM, Pracyk JB, Kirby M, Irani K, Goldschmith-Clermont PJ, Finkel T. Rac1 is required for cell proliferation and G2/M progression. Biochem J 1997;326:17–20.

[16] Yang XM, MacKenzie RE. NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase is the mammalian homolog of the mitochondrial enzyme encoded by the yeast MIS1 gene. Biochemistry 1993;32:11118–23.

[17] Peri KG, MacKenzie RE. NAD(+)-dependent methylenetetrahydrofolate dehydrogenase-cyclohydrolase: detection of the mRNA in normal murine tissues and transcriptional regulation of the gene in cell lines. Biochim Biophys Acta 1993;1171:281–7.

[18] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: encyclopedia for genes, proteins and diseases, Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot (Israel), 1997. World Wide Web URL: http://bioinfo.weizmann.ac.il/cards.