

Toxicogenomics screening of small molecules using high-density, nanocapillary real-time PCR

LAURA VASS¹, JÁNOS Z. KELEMEN¹, LILIÁNA Z. FEHÉR², ZSOLT LORINCZ³, SÁNDOR KULIN², SÁNDOR CSEH³, GYÖRGY DORMÁN⁴ and LÁSZLÓ G. PUSKÁS^{1,2}

¹Laboratory of Functional Genomics, Biological Research Center of the Hungarian Academy of Sciences, Temesvári krt. 62., 6726 Szeged; ²Avidin Ltd., Közép fasor 52., 6726 Szeged; ³TargetEx Ltd., Kápolna köz 4/a., 2120 Dunakeszi; ⁴Innobios LP, Kondorosi út 80., 1119 Budapest, Hungary

Received July 21, 2008; Accepted August 25, 2008

DOI: 10.3892/ijmm_00000102

Abstract. Compounds which induce toxicity through similar mechanisms lead to characteristic gene expression patterns. The concept that structurally similar compounds may have similar biological profiles, the so-called generalized neighborhood behavior, is less obvious to be demonstrated. We screened 625 compounds from a fully combinatorial library for their gene expression profiles *in vitro* over a selected toxicity panel of 56 genes. We used the novel nanocapillary, quantitative real-time PCR OpenArray™ technology that is coupling outstanding analytical performance with the medium-throughput ideal for such a sample-per-feature ratio. Applying a hybrid clustering on the gene expression data, correlation was analyzed between molecular scaffold and biological fingerprint. Structurally highly dissimilar, but similarly hepatotoxic compounds show similar fingerprint on our toxicity panel, however compounds of the same scaffold and of unknown biological effect do not always share similar fingerprints. Out of 12 different scaffolds, 4 families show non-correlating, uniform distribution among clusters whilst 8 families show neighborhood behavior of varying strength. Structurally not similar compounds may have highly similar biological activity, on the other hand, compounds of the same scaffold family do not all share the same biological effects based on toxicology related gene expression fingerprint.

Introduction

One of the biggest hurdles in drug development is the late-stage attrition caused by toxicity of drugs. Established drugs, such as Vioxx (Rofecoxib), Duract (Bromfenac), Pondimin

(Fenfluramine) have been withdrawn from the market because of unforeseen human toxicity. Early evaluation of drug safety during development through gathering predictive and causative information about potential toxicity could significantly reduce time, the later phase attrition and thus the overall expense of drug development (1-4).

Toxicogenomics (5) is a scientific field that studies how the genome is involved in responses to environmental stressors, toxicants and in general xenobiotics. Toxicogenomics combines studies of genomics, cell and tissue-wide protein expression and metabolomics to understand the role of gene-environment interactions in healthy and diseased samples.

Since the field of the '-omics' research is still establishing its own standards for both the preparative and the statistical evaluation phases, its application in toxicology studies of many variables and costly decisions is yet to be fully validated and accepted. Medicinal chemists and bioinformaticians have to closely cooperate and explore the confidence levels and limits of the techniques applied.

It is believed that compounds which induce toxicity through similar mechanisms lead to characteristic gene expression patterns (6). By clustering the gene expression profiles of well-characterized reference compounds and correlating these changes to standard toxicity indices, a gene expression fingerprint related to specific tissue or organ toxicity could be determined and applied to predict the toxicity of a candidate drug (5,6).

The concept of generalized neighborhood behavior that structurally similar compounds may have similar biological profiles have been demonstrated by a limited amount of studies on multiassay high-throughput screening data analysis (7-9). Yan *et al* (10) performed such analysis over the database of the Genomics Institute of the Novartis Research Foundation. Their objective was to identify results that are due to technology-related artifacts and target family specific activities. They also demonstrated the generalized neighborhood behavior using an initial database on 33107 compounds over 74 assays, with the strong emphasis on the fact that not all the structurally similar compounds can be expected to share the same biological activity profile.

In order to prove that different mechanisms of toxicity can be determined from gene expression fingerprints, Waring

Correspondence to: Dr László G. Puskás, Avidin Ltd., Közép fasor 52., 6726 Szeged, Hungary
E-mail: laszlo@avidinbiotech.com

Key words: toxicogenomics, cytotoxicity, real-time PCR, gene expression

et al (11) have treated rats with 15 known hepatotoxins and the gene expression data obtained by using microarray technology was clustered and correlated to histopathology and clinical chemistry results, which corresponded well.

Hamadeh *et al* (12,13) have tested *in vivo* the hypothesis that cDNA microarrays are an applicable platform for chemical-specific gene-expression profiling which profiles are characteristic across and within the structurally unrelated compound classes.

Van Delft *et al* (14,15) and colleagues investigated 20 chemical carcinogens over 597 toxicologically relevant genes via microarrays to prove the profoundly distinct gene expression patterns between genotoxic and non-genotoxic chemicals, applicable for classification.

Toxicogenomics can also characterize certain populations where the drug candidates could cause safety problems. Thus, toxicogenomics can not only be applied to provide an alert for potential toxicity but could differentiate between patient groups based on their responses and will play an important role in the development of personalized medicines.

Developments in biomarker research and in classification algorithms for genomics data support the establishment of large toxicogenomics databases, in fact more and more pharmaceutical companies have started to build their own database in hopes of predicting the potential toxicity of compounds and identify clearly which patient group is subjected by adverse effects. Publicly available databases, like the Comparative Toxicogenomics Database are also a result of this movement, and pharmaceutical companies are predicted to shift towards sharing their data and thus their costs.

We screened 625 compounds from a fully combinatorial library for their gene expression profiles *in vitro*, over 56 selected biomarkers. Our objective was to see to what extent their highly similar chemical structures induce similarities in their hepatotoxic fingerprints and to test the analytical performance of the nanocapillary, quantitative real-time PCR (QRT-PCR) technique and its general applicability for the field of toxicogenomics.

Preliminary tests have been performed with our inhouse ToxicoScreen DNA-microarrays (16) and with the traditional QRT-PCR technique, following which we shifted to the OpenArray nanocapillary QRT-PCR-technology (17) that has meanwhile appeared on the market (BioTrove Inc., Boston, MA, USA). This later technology merges the high-throughput of DNA-microarrays with the sound characteristics of QRT-PCR, therefore ideal for toxicogenomics screening of chemical libraries.

After determining the gene expression pattern of each compound we analyzed the results using hierarchical and K-means clustering methods looking for correlation between chemical scaffold and biological effect.

Materials and methods

Compounds. Using in-house validated chemical reactions that are suitable for parallel synthesis and a collection of multifunctional 'drug-like' building blocks, a dedicated discovery screening library of 10000 compounds has been enumerated by a cascading diversity building approach. For synthesis of the generated library high-throughput parallel

synthesis technology was applied which combines conventional medicinal chemistry practices with robot-assisted high-throughput techniques (18). In this matrix technology the intermediates are divided into small portions after each diversity-building synthetic step and reacted with a pool of different reagents in a parallel manner in isolated vessels. The generated non-exclusive library consists of 18 sub-libraries with an average 556 members in each (minimum 87, maximum 1461). As part of a large open access library of ~200000 structures, it has been tested and ranked for its medicinal chemistry attractiveness (19) and for its uniqueness (20).

Based on the cytotoxicity measured in an MRC-5 human fibroblast assay and the interpolated LD₅₀ values (from 6 concentrations using triplicates, $z' > 0.4$ for all plates), we selected 3x500 compounds (toxic $< 5 \mu\text{M}$, $5 \mu\text{M} <$ medium toxic $< 100 \mu\text{M}$, $100 \mu\text{M} <$ non-toxic) aiming maximal diversity within each group. Cytotoxicity of the selected 1500 compounds were measured in a HepG2 human hepatocarcinoma assay and out of these 625 structures were selected, keeping the desirable maximal diversity in view. Table I shows the scaffold structures of the major sublibraries (those with < 5 compounds not indicated) and the number of compounds generated with each scaffold. These 625 compounds were synthesized on a higher scale for gene expression analysis.

We used 8 commercially available compounds (pharmaceutical entities, pesticides) and 4 compounds presently under development as anti-cancer drugs for preparing positive control samples. These 12 compounds are listed in Table II. The samples were prepared the exact same way, only we applied several different concentrations of the compounds.

Cell treatment. HepG2 cells (European Collection of Cell Cultures, ECACC, Salisbury, UK) (3×10^5 cells/plate) were cultivated in DMEM medium (2% FBS) (Sigma-Aldrich Co., St. Louis, MO, USA) at 37°C in a humidified 5% CO₂ incubator. After 1 day of incubation, different toxic compounds (each compound was used approximately at one quarter of the measured cytotoxic LD₅₀ value) were added to the culture medium in DMSO (final solvent concentration 1%). We used an updated Beckman Biomek 2000 workstation for liquid handling. We used 200 μl tips from MBP (Molecular BioProducts Inc., San Diego, CA, USA). For the fluorescence measurement we used a VICTOR2 1420 Multilabel Counter from Perkin-Elmer (Perkin-Elmer Inc., Waltham, MA, USA). After 12 h incubation, cells were harvested and washed with PBS, 200 μl of RA1 (Macherey-Nagel GmbH & Co., Düren, Germany) was added containing 1% of β -mercaptoethanol. Cells were stored at -80°C until RNA purification procedure.

Cytotoxicity assay. Cytotoxicity was determined by using the AlamarBlue method (21) (Invitrogen Co., Carlsbad, CA, USA). During the assay optimization, we adjusted the following parameters: plating cell density, incubation time with AlamarBlue, medium composition, and tolerance of the compounds' solvents. HepG2 cells were cultivated at 37°C under 5% of CO₂ and 100% humidity. We used DMEM medium supplemented with 10% FCS (Sigma-Aldrich), and penicillin-streptomycin antibiotics. The initial cell number was 10^5 /well. During the assay, we reduced the FBS content

Table I. Markush structure and number of screened compounds for each scaffold [1-12].

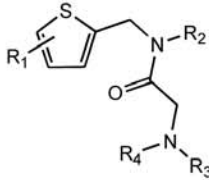
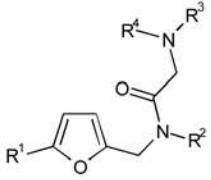
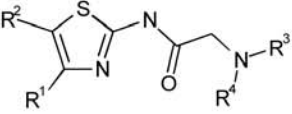
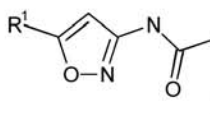
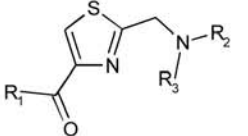
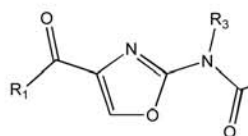
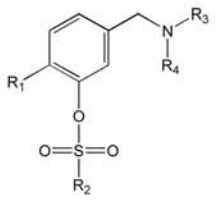
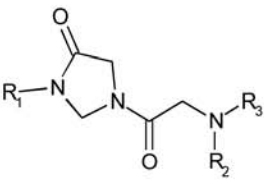
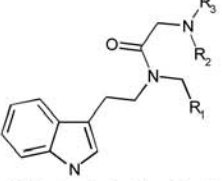
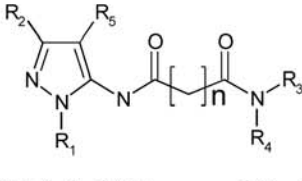
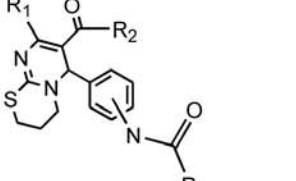
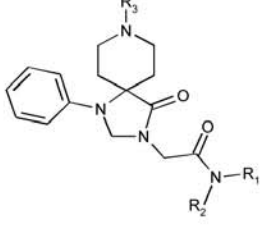
ID	Scaffold Markush structure	Comp/ Scaff	ID	Scaffold Markush structure	Comp/ Scaff
1	 N-Thiophen-2-ylmethyl-alkanamide	94	2	 N-Furan-2-ylmethyl-alkanamide	125
3	 N-Thiazol-2-yl-alkanamide	7	4	 N-Isoxazol-3-yl-alkanamide	4
5	 2-Aminomethyl-thiazole-4-carboxamide	9	6	 2-(Acylamino-methyl)-oxazole-4-carboxamide	7
7	 7a: R ₁ : H/Methoxy, R ₃ : Alkyl/Aryl/Hetaryl 7b: R ₁ is H, R ₃ : Alkyl Alkanesulfonic acid 4-aminomethyl-phenylester	26	8	 1-(2-Amino-acetyl)-imidazolidin-4-one	26
9	 N-[2-(1H-Indol-3-yl)-ethyl]-alkanamide	177	10	 N-(2-Methyl-2H-pyrazol-3-yl)-alkanamide	10
11	 6-(4-Acetylamino-phenyl)-8-methyl-3,4-dihydro-2H,6H-pyrimido[2,1b][1,3]thiazine-7-carboxamide	45	12	 2-(4-Oxo-1-phenyl-1,3,8-triazaspiro[4.5]dec-3-yl)-acetamide	95

Table II. Compounds used as positive control samples.

Name/ID	Application
K134	
K138	
Acetochlor	Pesticides
Dimetachlor	
Doxorubicin	Chemotherapy drug
Ivermectine	Anti-parasite medication
Sulfasalazine	Discontinued, drug for <i>colitis ulcerosa</i>
β -estradiol	Sex hormone
Ac201	
Ac202	Potential anti-cancer drugs under
Ac203	development
Ac204	

of the medium to 2% to avoid the masking effect on toxicity by FBS. We used the final volume of 80 μ l in a 384-well plate. We incubated the cells for 24 h with the compounds at various concentrations in the CO₂ thermostat. The concentration range of the compound depended on the compound's toxicity behavior. We used 1% concentration of the solvent (DMSO) and Triton X-100 (Merck & Co., Inc., NJ, USA) as a positive control. After 4 h of incubation with AlamarBlue, the fluorescence was measured by exciting at 544 nm and measuring emission at 590 nm. We calculated the cell viability and z' values. In all cases, the z' values were above 0.6. The LD₅₀ values were measured in six concentrations using triplicates.

Sample preparation. We isolated the RNA from the cell-samples using the ZR-96 Mini RNA Isolation Kit (Zymo Research Corp., Orange, CA, USA) filters and the NucleoSpin RNA elution liquids (Macherey-Nagel), with a DNase-treatment (NucleoSpin DNase I Set, Macherey-Nagel) inserted after desalting. The quantity and quality of RNA was assessed spectrophotometrically by a NanoDrop instrument (Rockland, DE, USA). Until cDNA conversion we stored samples at -80°C in the presence of 1 U/ μ l Prime Rnase inhibitor (Eppendorf AG, Hamburg, Germany). The isolated RNA was brought to the required 250 ng/ μ l concentration by lyophilization, then randomly converted to primed first strand cDNA, using a High Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA, USA). To reduce non-specific product formation during qPCR, the cDNA samples were heated to 75°C for 10 min to inactivate the reverse transcriptase; snap chilled for 5 min then treated with 1.3 U/ μ l Exonuclease I (USB Europe GmbH, Germany) for 1 h. The Exonuclease I was inactivated at 85°C for 10 min. The cDNA samples were stored at -20°C.

QRT-PCR. The PCR master mix consists of 1X LightCycler FastStart DNA Master SYBR Green I (Roche Applied Science, Indianapolis, IN, USA), 0.2% (w/v) Pluronic F-68 (Gibco, Carlsbad, CA, USA), 1 mg/ml BSA (Sigma-Aldrich), 1:4000 SYBR Green I (Sigma-Aldrich), 0.5% (v/v) glycerol

(Sigma-Aldrich), 8% (v/v) formamide (Sigma-Aldrich), MgCl₂ solution (Roche Diagnostics GmbH, Mannheim, Germany) and nuclease-free PCR-grade H₂O (Eppendorf) and sample. To test 56 genes, 4.5 μ l of reaction mix was prepared from each sample which were subsequently loaded onto the PCR-arrays.

Primers for the PCR reactions were designed using the program 'Primer Express' setting the following criteria: primer length 19-30, amplicon length maximum 150 nucleic acids, design for SYBRGreen reactions.

The PCR array thermal cycling protocol consisted of 10 min, 92°C polymerase activation step followed by 35 cycles of 15 sec at 92°C, 1 min at 55°C and 1 min at 72°C (imaging step). Following amplification, amplicon dissociation was measured by cooling the PCR array to 65°C then slowly heated to 92°C at 1°/min, with images collected every 0.25°C. The PCR experiments were done using the nanocapillary QRT-PCR instrument developed at BioTrove Inc.

Statistics. We performed the following statistical evaluation steps on the raw output data files for each open-array plate: i) We transformed the data matrices (48* 8x8) and normalized the Ct-values of the individual PCR runs (each sample) for the average values of the housekeeping genes on each sub-array (8x8, Δ Ct) with our in-house developed software, thus making these transformations automated, eliminating the plausible human error. All values are in logarithmic values of base 2, as inherent to the PCR technique. Data for cycles where the Ct-values are above 28 and Δ Ct-values are above 10 were eliminated from the evaluation. ii) In case too few data points had been obtained for a gene throughout all samples (i.e., <20% gave acceptable Δ Ct), that gene was excluded from further analysis. Similarly, samples that had not given an acceptable Δ Ct throughout >20% of all 56 genes were excluded. iii) Having the raw data so transformed, we merged all into one database. The average expression level (cycle number) for all negative control (i.e., only vehicle-treated) samples was calculated over each gene. This average was subtracted from each sample's expression level over each gene, resulting in the $\Delta\Delta$ Ct values. Missing data points are omitted from clustering. This database is available online at <http://www.brc.hu/~chiplab/toxicogenomics/data.txt>. iv) We applied a hybrid clustering method: unsupervised hierarchical clustering methods with uncentered correlation similarity metrics, average linkage clustering, then based on the achieved clusters, supervised k-means and k-medoids clustering (for detailed description of these statistical methods see 22,23). Clustering was performed by the software 'Cluster', output from the clustering was visualized by the software 'TreeView', both programs developed for statistical organization and graphical display of microarray data by Eisen *et al.* (24). Missing data points are omitted by the clustering algorithms. Because the resulting clusters depend on the initial random assignments, it is a common practice to run the clustering algorithm several times and return the best clustering found. v) Based on the scaffold structure or the characteristic residues, we assigned the tested chemicals into subgroups (hereinafter referred to as scaffold-libraries). vi) We performed Pearson's χ^2 test on the obtained clusters, using the statistical program 'R' (<http://www.r-project.org>).

Table III. The discovery gene set of 56 toxicology marker genes ('Tox-I' OpenArray Plate, Avidin Ltd.).

No.	Gene product	Accession no.
1	GDF15 growth differentiation factor 15	NM 004864
2	SOD1 superoxide dismutase 1	NM 000454
3	PRDX1 peroxiredoxin 1	NM 002574
4	UDP-glucose dehydrogenase UGDH	NM 003359
5	HSPE1 heat shock 10 kDa protein 1	NM 002157
6	EPHX1 epoxide hydrolase 1	NM 000120
7	PPIA peptidylprolyl isomerase A (cyclophilin A)	NM 021130
8	GSTP1 glutathione-S-transferases	NM 000852
9	HSPCA heat shock 90 kDa protein 1 α	NM 005348
10	HSPA1A heat shock 70 kDa protein 1A	NM 005345
11	CAT catalase	NM 001752
12	RAD50 homolog	NM 005732
13	CYP20A1 cytochrome P450 monooxygenase	NM 020674
14	GLUL glutamate-ammonia ligase	NM 001033044
15	PPARA peroxisome proliferative activated receptor α	NM 005036
16	CPT1A carnitine palmitoyltransferase	NM 001031847
17	TPMT thiopurine S-methyltransferase	NM 000367
18	GSTT1 glutathione S-transferase θ 1	NM 000853
19	RPLP0 ribosomal protein large P0	NM 001002
20	PRDX2 peroxiredoxin 2	NM 181737
21	TP 53 tumor protein p53	NM 000546
22	NQO1 NAD(P)H dehydrogenase quinone 1	NM 001025434
23	GADD45A growth arrest, DNA-damage-inducible α	NM 001924
24	GSR glutathione reductase	NM 000637
25	FTL ferritin light polypeptide	NM 000146
26	LTA4H leukotriene A4 hydrolase	NM 000895
27	HOX1 heme oxygenase 1	NM 002133
28	GPX1 glutathione peroxidase 1	NM 000581
29	PPARG peroxisome proliferative activated receptor γ	NM 138711
30	COMT catechol-O-methyl transferase	NM 000754
31	POR cytochrome P450 reductase	NM 000941
32	PCNA proliferating cell nuclear antigen	NM 002592
33	GPX4 glutathione peroxidase 4	NM 002085
34	SOD2 superoxide dismutase 2 mitochondrial	NM 000636
35	CYP1B1 cytochrome P450 1 B1	NM 000104
36	IL6 interleukin 6	NM 000600
37	IGFBP6 insulin-like growth factor binding protein 6	NM 002178
38	OAT ornithine aminotransferase	NM 000274
39	GSTM3 glutathione S-transferase M3	NM 000849
40	CASP3 caspase 3 apoptosis-related cysteine peptidase	NM 032991
41	PGK1 phosphoglycerate kinase 1	NM 000291
42	CYP1A1 cytochrome P450 1 A1	NM 000499
43	CYP1A2 cytochrome P450 1 A2	NM 000761
44	CYP2C9 cytochrome P450 2 C9	NM 000771
45	CYP2E1 cytochrome P450 2 E1	NM 000773
46	CYP3A5 cytochrome P450 3 A5	NM 000777
47	CYP19A1 cytochrome P450 19 A1	NM 031226
48	UGT1A4 UDP glycosyltransferase 1 A4	NM 007120
49	MT2A metallothionein 2A	NM 005953
50	MT3 metallothionein 3	NM 005954
51	GPX2 glutathione peroxidase 2	NM 002083
52	FMO3 flavin-containing monooxygenase 3	NM 001002294
53	NOS2A nitric oxide synthase 2A	NM 000625
54	RBP4 retinol binding protein	NM 006744
55	IL1B interleukin 1 β	NM 000576
56	CRYAB crystallin α B	NM 001885

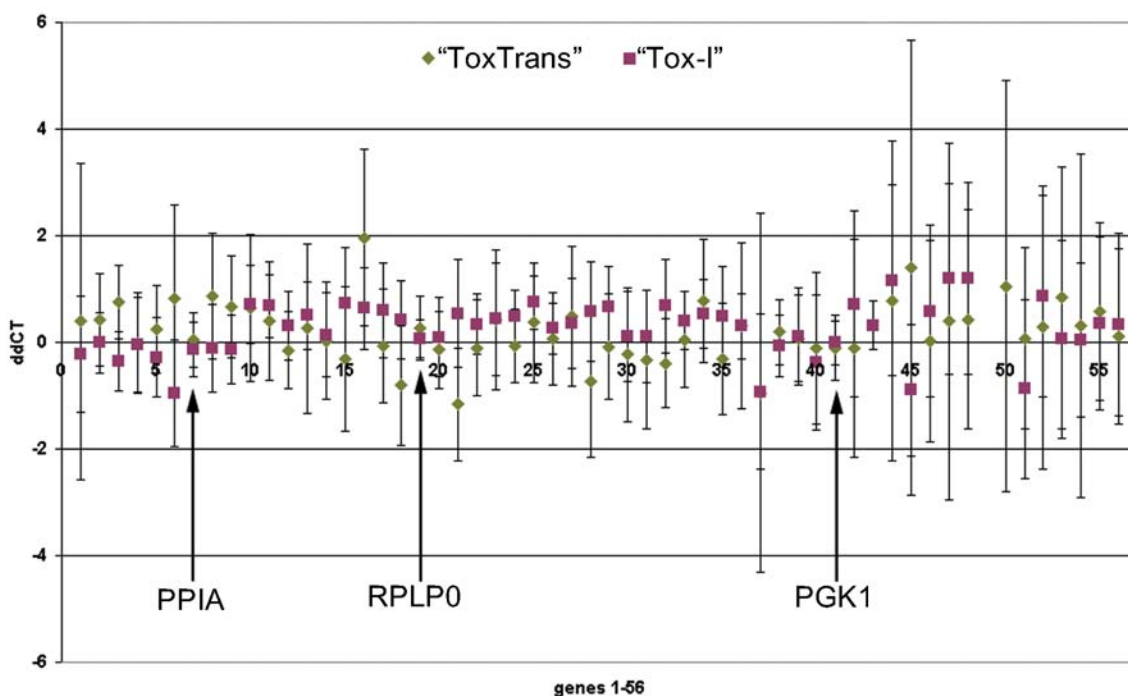


Figure 1. Mean values and standard deviation for all compounds tested on OpenArray plate batches 'AQY' and 'ATX', over all evaluated genes. The 3 indicated genes, PPIA peptidylprolyl isomerase A (cyclophilin A, gene 7), RPLP0 ribosomal protein large P0 (gene 19), PGK1 phosphoglycerate kinase 1 (gene 41) and ribosomal protein large P0 are of significantly low deviation. As a proof-of-concept, this underlines using these genes as housekeeping genes for normalization of the Ct values.

However, correlation statistics are only to be used as guidance in our case, since our sample \times gene clusters contain <100 data points each. As written by Simon *et al.* (25), 'clustering is a subjective technique, whose results are highly influenced by selection of the clustering algorithm and similarity metric'. Since we had no preliminary expectations about the gene expression pattern over the discriminatory gene set for the tested combinatorial library, based on the correlation coefficient for scaffold vs. expression-level cluster, we have changed the parameters in steps 4 and 5 to obtain the most distinct clusters.

Results

To assess gene expression changes in human hepatic cells in response to different cytotoxic compounds, we used the nanoliter, high-throughput QPCR technology (17). The technology developed at BioTrove Inc. is a hybrid approach for performing QRT-PCR in an array of 3072 isolated nanoscale through-holes. Up to 48 different cDNA samples can be tested, with 64 separate reactions (out of which 56 can be custom-chosen genes and there are 8 control-reactions) in each of the 48 subarrays on one plate; with the thermal cycler handling 3 plates, up to 9216 qPCR reactions can run in about 4 h. The 'Tox-I' OpenArray plate was designed by Avidin (Avidin Ltd., Szeged, Hungary). The discovery gene set used in this study is listed in Table III.

We performed an initial testing for analytical performance of the nanocapillary QRT-PCR instrument with a separate batch of OpenArray Plates coded 'TransTox', containing both the said 56 toxicology marker gene primers and 56 other primers for transporter genes (www.avidinbiotech.com), in

total 112 genes. For this initial testing, 60 different samples, including 36 out of the selected 625 small, drug-like compounds of unknown effects from the combinatorial library, 12 commercially available toxic compounds (pharmaceutical entities, pesticides) of known effects and of yet unknown effects and 8 proprietary anti-cancer drugs presently under development (Avidin Ltd.) as positive controls and vehicle-treated negative control samples were used. A total of 119 reactions were evaluated for average values of expression for each gene and the relevant standard deviation.

Results are presented in Fig. 1. for the toxicology gene markers and the housekeeping genes, the later indicated by arrows. Reactions for the gene MT2A metallothionein 2A (gene 49) did not work thus this gene was eliminated from further evaluation. The very low deviation of the $\Delta\Delta Ct$ values for the genes specially indicated (cyclophilin A, phosphoglycerate kinase 1, ribosomal protein large P0) underline their use as housekeeping genes, especially when compared to the wide-scale changes over the full discovery gene set.

The same statistical evaluation was performed for the 'Tox-I' coded plate-batch (only the 3 housekeeping genes and the toxicity gene markers) in which we tested all 625 compounds from the combinatorial library along with 12 of the above mentioned positive control samples (Table II) and at least one vehicle-treated negative control sample was running on each OpenArray plate, with similar results.

Fig. 2 is the graphical presentation of the clustering results for the samples. In case too few data points had been obtained for a gene throughout all samples (i.e., $<20\%$ gave acceptable ΔCt -values), that gene was excluded from further analysis. Similarly, samples that had not given an acceptable ΔCt -value throughout $>20\%$ of all 56 genes were excluded.

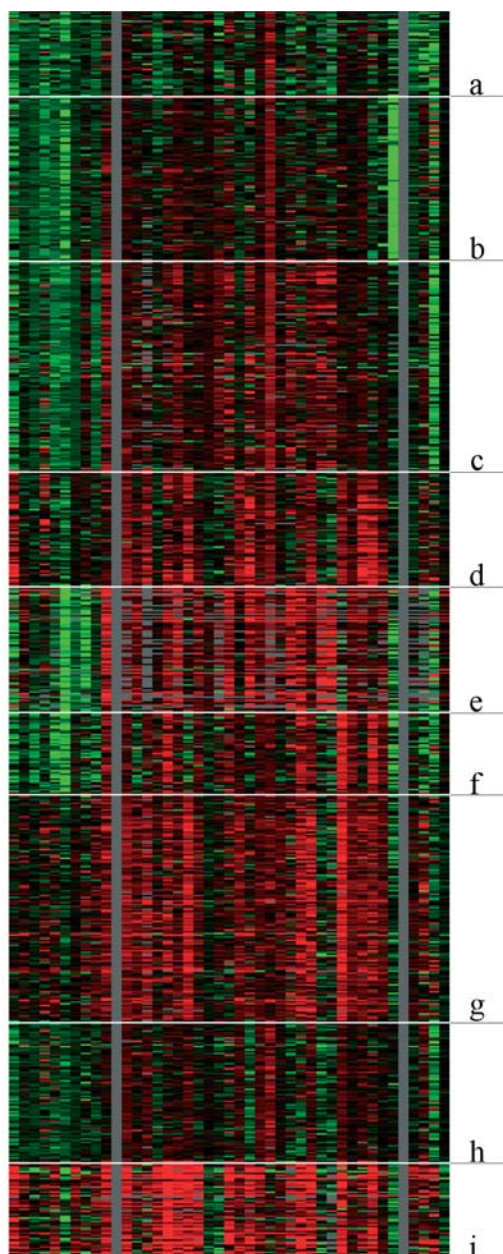


Figure 2. Graphical presentation of the clustering results for the samples, wherein each column is representing values for a gene, and each row is representing a sample. Green values are for repression, red values are for overexpression. Grey fields mark missing data, these missing values were not considered for clustering. Using the supervised K-means clustering method, samples were ordered into nine clusters, genes into three.

We applied K-means clustering method with nine nodes for samples and three nodes for genes. These numbers were deduced from the results of unsupervised clustering. Scaffold families of only one or two members have been excluded from correlation analysis.

Fig. 3 illustrates the results for correlation between the scaffold-families and the gene expression changes generated by compounds belonging to them. In the figure, values are the number of compounds belonging to each scaffold family [1-12] in each cluster [a-i], divided by the number of compounds in the given scaffold family and by the number of compounds in the given cluster. Thus the values are normalized for both the size of the given scaffold-library and the size of the given cluster. The values are not comparable across scaffold types. For better understanding, uniform distribution over clusters of a family would mean no correlation, that is no neighborhood behavior, but the less uniform the distribution the more correlation there is between structure and biological activity.

Discussion

We screened 625 compounds from a fully combinatorial library for their gene expression profiles *in vitro* in HepG2 cells, over a discovery gene set of 56 selected biomarkers. The scaffolds of these compounds are relatively similar, containing 5-6 membered (aromatic) ring(s) that may contain N, O or S as heteroatom(s). The libraries selected for each scaffold differ very broadly in size, from 4 compounds up to 177 compounds per scaffold.

By the combination of a relatively big combinatorial chemical library and a relatively small set of selected toxicological biomarkers, we intended to avoid the two culprits of toxicogenomics: 'the curse of dimensionality' (too many genes), and 'the curse of dataset sparsity' (too few samples). The generally accepted, however rarely adapted sample-per-feature ratio for robust clustering performance is at least 5-10 (26). In the present experiment, this number is approximately 12.

The statistical evaluation of the results was aimed at determining whether there is strong - if any - neighborhood behavior among samples of the same scaffold based on their hepatotoxic fingerprints, as well as testing the analytical performance and applicability of the applied nanocapillary QRT-PCR technique for measuring such correlation.

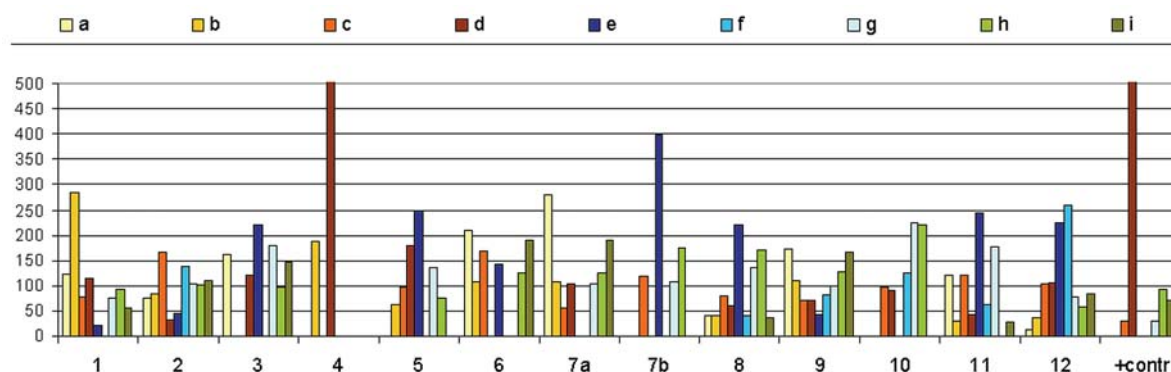


Figure 3. The number of each scaffold-family in each cluster, normalized for both the size of the given scaffold-library and the size of the given cluster.

Table IV. The number of samples from compounds of each scaffold family [1-12] distributed in each cluster [a-i] and the summed total of samples for each scaffold.

	1	2	3	4	5	6	7a	7b	8	9	10	11	12	+control
<i>a</i>	8	7	1	0	0	1	4	0	1	22	0	4	1	0
<i>b</i>	36	15	0	1	1	1	3	0	2	27	0	2	6	0
<i>c</i>	13	37	0	0	2	2	2	1	5	22	2	10	22	1
<i>d</i>	10	4	1	3	2	0	2	0	2	12	1	2	12	14
<i>e</i>	2	6	2	0	3	1	0	2	8	8	0	12	28	0
<i>f</i>	0	12	0	0	0	0	0	0	1	10	1	2	21	0
<i>g</i>	13	25	3	0	3	0	4	1	9	34	5	16	18	1
<i>h</i>	10	15	1	0	1	1	3	1	7	27	3	0	8	2
<i>i</i>	4	11	1	0	0	1	3	0	1	23	0	1	8	1
Σ	96	132	9	4	12	7	21	5	36	185	12	49	124	19

In recent years there has been many conceptual developments in the field of the miniaturization of PCR devices (27). In case of a large combinatorial library there is only a very limited amount of each compound, thus we need a high-throughput platform with stringent diagnostic standards. One of the set-backs to be overcome when working with low sample volumes is the fast evaporative loss along thermal cycling, moreover the increasing surface-to-volume ratio brings biochemical surface absorption problems along.

The OpenArray Cyler from BioTrove Inc. overcomes both problems. The system joins high accuracy, precision and dynamic range characteristics of QRT-PCR with the relatively higher throughput of microarrays (28). Up to 3072 individual solution-phase reactions are run in parallel in 33 nl through-holes on the size of a microscope slide in the software-controlled, completely standardized environment of a thermal cycler. The custom-selected primer pairs are immobilized in the OpenArray plate generating a custom-based screening platform for a subset of genes. This technology is ideal for toxicogenomics screening. In comparison to microarrays, a higher analytical performance is due to a more standardized and automatized loading and incubation of samples. Cross-contamination of samples is both theoretically and provenly eliminated on the plates. With the given 56 genes, throughput is higher and a smaller amount of the investigated chemicals are necessary for the incubation with the cell cultures to give the sufficient RNA-quantity. In case of a combinatorial chemical library, this is an important factor. From all these advantages comes the lower cost per sample. The technology does not have the usual 'gradient-problem' of microarrays which was also very important in our case; changes in expression levels of tests *in vitro* are often close to the background noise level of microarrays, and false negatives or positives are clearly dangerous when developing a fingerprint-analytical assay.

On a conventional real-time PCR instrument, the analysis of 700 samples for 60 genes would have required roughly 42000 individual reactions. In one PCR run we analysed up to 384 reactions, which in case of SYBRGreen detection includes a housekeeping genes on each tested samples. Thus,

such a study would result in roughly 110 PCR runs, that is about 240 h of runtime, not including sample preparation or loading. In case of microarrays, the throughput of samples would be even lower, however the gene-set screened would be several order of magnitudes higher. The cost of such a study would be beyond the scope of this project, moreover by using the nanocapillary PCR system we could avoid the so-called curse of dataset sparsity and of dimensionality.

The results in Fig. 3 and the adjacent Table IV are to illustrate correlation between the scaffolds and the gene expression changes the compounds with a given scaffold induce. The values indicated are the number of compounds with a certain scaffold [1-12] in each [a-i] sample-cluster, as well as the distribution of the positive control samples over the same clusters, normalized to the size of the clusters and to the size of the scaffold-library. Thus the normal distribution here would be the uniform distribution. The less uniform the distribution of a scaffold-library over the clusters a-i, the more correlation there is between scaffold and gene expression levels for these toxicity markers, i.e., the stronger the neighborhood behavior of the given scaffold. This statistical evaluation of a library for the correlation between scaffold structure and the induced gene expression levels is more robust for larger libraries. However, we have indicated also the smaller libraries, because the distribution can be meaningful as well, for instance in case of scaffold-library no. 4. Some scaffold-types, such as type no. 1 or type no. 4 are showing strong neighborhood behavior, whilst others such as type no. 2 or type no. 9 do not show much correlation. Table IV contains the number of compounds in a scaffold vs. cluster representation, hence the 14x9 matrix.

From the statistics point of view, the scaffold libraries no. 12 (124 samples), no. 2 (132 samples) and no. 9 (185 samples) are of most interest. Compounds of the structure no. 12, that is the 2-(4-Oxo-1-phenyl-1,3,8-triaza-spiro[4.5]dec-3-yl)-acetamide scaffolds are most prevalent in clusters *e* and *f*. In the *e* cluster, genes for EPHX1 and GSTP1 are showing strong repression, in case of some samples the IGFBP6 gene is repressed, whilst transcription of the following are induced: HSPA1A, CPT1A, TP53, GADD45A,

Table V. Relative scaffold representation in sample clusters.

Scaffold type	(%) in <i>a</i>	(%) in <i>b</i>	(%) in <i>c</i>	(%) in <i>d</i>	(%) in <i>e</i>	(%) in <i>f</i>	(%) in <i>g</i>	(%) in <i>h</i>	(%) in <i>i</i>	Total no. of scaffold (100%)
1	8.3	37.5	13.5	<i>10.4</i>	2.1	0.0	13.5	10.4	4.2	96.0
2	5.3	11.4	28.0	3.0	4.5	9.1	18.9	11.4	8.3	132.0
7	19.0	14.3	9.5	9.5	0.0	0.0	19.0	14.3	<i>14.3</i>	21.0
8	2.8	5.6	13.9	5.6	22.2	2.8	25.0	<i>19.4</i>	2.8	36.0
9	11.9	14.6	11.9	6.5	4.3	5.4	18.4	14.6	12.4	185.0
11	8.2	4.1	20.4	4.1	<i>24.5</i>	4.1	32.7	0.0	2.0	49.0
12	0.8	4.8	17.7	9.7	22.6	<i>16.9</i>	14.5	6.5	6.5	124.0

In italics are the highest scaffold representation in a particular sample cluster, the bold figures indicate the highest number of compounds within the scaffold classes which belong to the same sample cluster, if both in italics and bold they belong to both.

HOX1, GPX1, COMT, POR. Cluster *f* is similar in the characteristic repression of the EPHX1 and the IGF1BP6 gene, but that of HSPE1 as well; the genes PPARA, PCNA, GPX1 and CYP1B1 are induced. Induction of CYP1B1 is most characteristic in clusters *d* and *i*. The libraries 8 (36 samples) and 11 (49 samples) show similar distribution to the library no. 12 and some of the smaller libraries as well [5, 3 and 7b]. Libraries no. 2 of the N-Furan-2-ylmethyl-alkanamide scaffolds and no. 9 of the N-[2-(1H-Indol-3-yl)-ethyl]-alkanamide scaffolds show similarities and are definitely distinct from those mentioned before. Type no. 7a (21 samples) scaffolds are mostly present in the clusters *a* and *i* which two show opposite tendencies for most genes. They are, however, absent from clusters *e* and *f* which are highly similar. The scaffolds type number 1 (96 samples) are most abundant in cluster *b*, this is most probably due to the significant repression of the IGF1BP6 and the EPHX1 genes. Alike scaffolds no. 7a, they are almost absent from clusters *e* and *f*. It is only scaffolds no. 4 (4 samples) and the positive control samples that are also absent from the *e* and *f* clusters. The positive control samples are most abundant in cluster *d*, showing strong co-induction of several genes: GDF15, UGDH, PPARA, CPT1A, TPMT, NAD(P)H, FTL, SOD2, CYP1B1 and slight repression of the gene EPHX1. Distribution of the compounds of scaffold type no. 10 (a total of 12 samples) and nos. 7a and 7b (21 and 5 compounds in total, respectively) is not clearly distinguishable from random distribution due to the low number of compounds. Scaffold type no. 10 compounds are mostly clustered in clusters *g* and *h*, showing general induction, mostly for the genes TPMT and PCNA.

By omitting scaffold clusters representing 20 compounds or less from the evaluation of the scaffold representation in the sample clusters, the following statistical distribution can be observed (Table V): in sample clusters *a*, *b* and *c* the highest scaffold representation correlated well with the highest number of compounds within the scaffold classes which belong to the same sample cluster (19, 38 and 28%, respectively). This clearly indicates that the sample cluster correlates well with the scaffold structure regardless their substitution pattern. Cluster *g* shows similar behavior (scaffold no. 11, 32%) except that there are another 2 scaffold

classes where compounds are highly represented in this sample cluster (25% of scaffold no. 8 and 18% of scaffold no. 9). Interestingly, these scaffolds are fairly unrelated. Scaffold no. 11 is also highly represented in sample cluster 'e' (24%) together with no. 12 (22%).

Sample cluster *d* is one of the most interesting clusters since most of the toxicology clusters belong to this cluster. Interestingly, only approximately 7% of all compounds belong to that sample cluster and the compounds were evenly distributed within the scaffold classes. The highest scaffold representation was 10% (scaffold class no. 1, thiophene). By analyzing the structure of the 44 compounds in cluster *d* (particularly the substitution pattern) and the distribution of their physicochemical parameters, no significant correlation was found.

Relatively even scaffold class distribution was observed in sample cluster *h*, with moderate representation within the whole compound library (11% of all compounds).

Sample cluster *i* is between the *a*, *b*, *c* as well as *d* sample clusters with 8% overall compound representation, with two major scaffold classes (nos. 7 and 9), however little structural relationship can be identified within these two scaffold classes.

From our hybrid clustering method applied on the data achieved as described above, nine clusters were formed from the tested samples. These clusters contain compounds of different scaffolds. The statistical evaluation of the distribution of these scaffolds over the gene expression clusters leads to the two following conclusions.

Structurally not similar compounds may have highly similar biological activity: cluster *b* for instance is a very tight, uniform cluster, however there is ten types of scaffolds in it. This is underlined by the experienced fingerprints of the applied positive control samples: there are structurally very different compounds, slight structural variations of which have been also tested even in different concentrations, yet they are mostly directed into the same cluster (cluster *d*) and compounds of the exact same structure but different incubation concentration are not always found to be most similar.

On the other hand, compounds of the same scaffold family do not all share the same biological effect. Compounds type no. 12 of the 2-(4-Oxo-1-phenyl-1,3,8-triaza-spiro[4.5]dec-3-

yl)-acetamide scaffold show a non-uniform distribution based on gene expression results, whilst the family no. 2 of the N-Furan-2-ylmethyl-alkanamide scaffold is almost uniformly distributed over the expression clusters. This difference however cannot be explained by their scaffolds. Yan *et al* (10) reached the same conclusion, even when starting from a significantly bigger database.

In case of testing a completely unknown library of chemical structures, without having preliminary information on the LD₅₀ values, compounds are usually applied at the same concentration. Gene expression data from the selected toxicity panel would correlate rather with toxicity than with the chemical scaffolds.

Clustering results for the same set of molecules over a different set of genes that are connected to the basic biological effects of these molecules - unlike our discovery gene set, giving us information on the indirect, toxic effects - would most probably be more obviously according to the molecular scaffolds.

For finding the correlation between a library of molecular scaffolds and their general biological fingerprint, one would perform prescreening over the full genome and with selected marker genes look for correlation patterns. The best markers for such analysis would most probably not be those measuring toxicity. For attaining information on the initial, toxic side effects of these scaffolds, one would however screen with a selected toxicity panel. As apparent from our study as well, with this later, selected toxicology gene set one does not expect stringent results for neighborhood behavior, but the results gained are more informative from a toxicologist's point of view.

Acknowledgements

This study was partly supported by the 'Ányos Jedlik - AVINOMID' and the 'Oszkár Asbóth' NKTH-XTTPSRT1 grants from the National Office for Research and Technology (NKTH) and GVOP-3.1.1-2004-05-0280/3.0 grant of the Hungarian Government. We are grateful to the dedicated chemists of AMRI (formerly ComGenex) for the preparation of the combinatorial library, from where the compounds were selected in the present study and for Miklós J. Szabó for his contribution in calculations.

References

1. Khor TO, Ibrahim S and Kong AN: Toxicogenomics in drug discovery and drug development: potential applications and future challenges. *Pharm Res* 23: 1659-1664, 2006.
2. Stevens JL: Future of toxicology - mechanisms of toxicity and drug safety: where do we go from here? *Chem Res Toxicol* 19: 1393-1401, 2006.
3. Fielden MR and Kolaja KL: The state-of-the-art in predictive toxicogenomics. *Curr Opin Drug Discov Devel* 9: 84-91, 2006.
4. Kramer JA, Sagartz JE and Morris DL: The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov* 6: 636-649, 2007.
5. Nuwaysir EF, Bittner M, Trent J, Barrett JC and Afshari CA: Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog* 24: 153-159, 1999.
6. Darvas F, Dormán G, Krajcsi P, *et al*: Recent advances in chemical genomics. *Curr Med Chem* 11: 3119-3145, 2004.
7. Weinstein JN, Myers TG, O'Connor PM, *et al*: An information-intensive approach to the molecular pharmacology of cancer. *Science* 275: 343-349, 1997.
8. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S and Weinstein JN: Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J* 2: 259-271, 2002.
9. Horvath D and Jeandenans C: Neighborhood behavior of *in silico* structural spaces with respect to *in vitro* activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* 43: 680-690, 2003.
10. Yan SF, King FJ, He Y, Caldwell JS and Zhou Y: Learning from the data: mining of large high-throughput screening databases. *J Chem Inf Model* 46: 2381-2395, 2006.
11. Waring JF, Jolly RA, Ciurlionis R, *et al*: Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol Appl Pharmacol* 175: 28-42, 2001.
12. Hamadeh HK, Bushel PR, Jayadev S, *et al*: Gene expression analysis reveals chemical-specific profiles. *Toxicol Sci* 67: 219-231, 2002.
13. Hamadeh HK, Bushel PR, Jayadev S, *et al*: Prediction of compound signature using high density gene expression profiling. *Toxicol Sci* 67: 232-240, 2002.
14. Van Delft JH, van Agen E, van Breda SG, Herwijnen MH, Staal YC and Kleinjans JC: Discrimination of genotoxic from non-genotoxic carcinogens by gene expression profiling. *Carcinogenesis* 25: 1265-1276, 2004.
15. Van Delft JH, van Agen E, van Breda SG, Herwijnen MH, Staal YC and Kleinjans JC: Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. *Mutat Res* 575: 17-33, 2005.
16. Vass L, Kis Z, Feher LZ, *et al*: Medium-throughput microarray-based approach for toxicogenomic profiling of small molecules. *QSAR Comb Sci* 25: 1039-1046, 2006.
17. Morrison T, Hurley J, Garcia J, *et al*: Nanoliter high throughput quantitative PCR. *Nucleic Acids Res* 34: E123, 2006.
18. Darvas F, Dorman G, Urge L, Szabo I, Ronai Z and Sasvari-Szekely M: Combinatorial chemistry. Facing the challenge of chemical genomics. *Pure Appl Chem* 73: 1487-1498, 2001.
19. Baurin N, Baker R, Richardson C, *et al*: Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J Chem Inf Comput Sci* 44: 643-651, 2004.
20. Verheij HJ: Leadlikeness and structural diversity of synthetic screening libraries. *Mol Divers* 10: 377-388, 2006.
21. Pagé B, Pagé M and Noel C: A new fluorimetric assay for cytotoxicity measurements *in vitro*. *Int J Oncol* 3: 473-479, 1993.
22. Maggioli J, Hoover A and Weng L: Toxicogenomic analysis methods for predictive toxicology. *J Pharmacol Toxicol Methods* 53: 31-37, 2006.
23. Svrakic NM, Nestic O, Dasu MRK, Herndon D and Perez-Polo JR: Statistical approach to DNA chip analysis. *Recent Prog Horm Res* 58: 75, 2003.
24. Eisen MB, Spellman PT, Brown PO and Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863-14868, 1998.
25. Simon R, Radmacher MD, Dobbin K and McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95: 14-18, 2003.
26. Somorjai RL, Dolenko B and Baumgartner R: Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19: 1484-1491, 2003.
27. Zhang C and Xing D: Miniaturized PCR chips for nucleic acid amplification and analysis: latest advances and future trends. *Nucleic Acids Res* 35: 4223-4237, 2007.
28. Stedtfeld RD, Baushke SW, Tourlousse DM, *et al*: Development and experimental validation of a predictive threshold cycle equation for quantification of virulence and marker genes by high-throughput nanoliter-volume PCR on the OpenArray platform. *Appl Environ Microbiol* 74: 3831-3838, 2008.