DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Computational Search for Possible Mechanisms of 4-Thiazolidinones Anticancer Activity: The Power of Visualization

Oleg Devinyak,[a] Dmytro Havrylyuk,[b] Borys Zimenkovsky[b] and Roman Lesyk*[b]

Abstract: Public databases of NCI-60 tumor cell line screen results and measurements of molecular targets in the NCI-60 panel give the opportunity to assign possible anticancer mechanism to compounds with positive outcome from antitumor assay. Here, the novel protocol of NCI databases mining where inferences are based on the visualization is presented and utilized with the aim to identify putative biological routes of 4-thiazolidinones anticancer effect. As a result, highly potent 4-thiazolidinone-pyrazoline-isatin conjugates show the similarity of activity patterns with puromycin and CBU-028 and their pattern is also highly correlated with fraction of methylated CpG sites in CD34, AF5q31 and SYK. Several compounds from this group show strong negative correlation with fraction of methylated CpG sites in HOXA5. Thiopyrano[2,3-*d*][1,3]thiazol-2-ones bearing naphtoquinone fragment were found to possess the same activity pattern as fusarubin does. But none of the studied 4-thiazolidinone derivatives has activity fingerprint similar to standard anticancer agents. The obtained results bring medicinal chemistry closer to the understanding of basic nature of 4-thiazolidinones effect on cancer cells.

Keywords: Activity patterns, Anticancer activity, Database mining, 4-Thiazolidinones, Visualization

1 Introduction

The search for new anticancer agents is one of the leading directions in medicinal chemistry. Developmental Therapeutics Program (DTP) of National Cancer Institute, USA (NCI) plays an important role in the discovery of antitumor drugs providing support both to private and academic researchers.[1] The NCI-60 Tumor Cell Line Screen is a part of the DTP anticancer drug discovery program and is designed to screen compounds for potential anticancer activity utilizing 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney.[2] This screen is unique in that the complexity of a 60 cell line dose response produced by a given compound results in a biological response pattern. Using the activity pattern it is possible to assign a putative mechanism of action to a test compound, or to determine that the response pattern is unique and not similar to that of any of the standard prototype compounds included in the NCI database. In addition, following characterization of various cellular molecular targets in the 60 cell lines, it may be possible to select compounds most likely to interact with a specific molecular target.[3,4] For the purpose of activity pattern analysis, the COMPARE program has developed and is accessible at http://dtp.nci.nih.gov/compare/.[5,6] More comprehensive web-based tool for NCI data analysis have appeared recently.[7] Consequently, multitarget approach to the search of new antitumor medicines was given much attention during the last years [8-10].

Over 700 4-thiazolidinone-based compounds from

[a] *Department of Pharmaceutical Disciplines, Uzhgorod National University*
Narodna sq. 1, 88000 Uzhgorod, Ukraine

[b] *Department of Pharmaceutical, Organic and Bioorganic Chemistry, Danylo Halytsky Lviv National Medical University*
Pekarska str. 69, 79010 Lviv, Ukraine
**e-mail: dr_r_lesyk@org.lviv.net, phone/fax: +38(032)275-5966/+38(032)275-7734*

Supporting Information for this article is available on the WWW under www.molinf.com our in-home library were tested since 2005 for cytotoxicity in vitro against NCI-60 cancer cell-line panel and significant number of compounds possessing different levels of anticancer activity have been identified.[11,12] The majority of papers reporting synthesis and biological evaluation of these 4-thiazolidinones were supported with COMPARE analysis of hit compounds anticancer pattern.[13-19] The results show that the correlation between activities of highly-potent 4-thiazolidinones and standard anticancer agents is rather weak, supposing some non-common or even novel mechanism of biological action. The closest to 4-thiazolidinones known patterns belong to tubulin polymerization inhibitors,[13,14] dihydroorotate dehydrogenase inhibitors,[15] topoisomerase II inhibitors and alkylating agents,[16] inhibitors of protein or RNA synthesis.[17] The diversity of these patterns assumes the presence of several different mechanisms in the 4-thiazolidinones library. But it remains unclear whether these mechanisms correspond to above mentioned or they present some new ones. Moreover, since only sporadic observations were available, there is a great chance to be deceived by chance correlation. This work provides systematic computational search for possible mechanisms and is intended to shed some light on the routes of 4-thiazolidinones anticancer activity. The methods used to discover the knowledge from the database of biological evaluation results belong to the field of data mining.

The idea of data mining in the NCI databases is not new. Shi et al. published a series of papers devoted to the cluster analysis of NCI-60 tumor cell line screen results.[20-22] They used hierarchical cluster analysis with average linkage to find relations between molecular structure of ellipticine analogs and their activity pattern and additionally have found correlation between p53 status of cell culture and its sensitivity to ellipticine derivatives.[20,21] Also, all public (as for 1999 year) compounds from NCI database were studied with principal component analysis and hierarchical clustering.[22] The relationships between gene expression patterns and anticancer activity patterns were investigated by Scherf et al. in 2000.[23] Fang and co-workers have developed web-based (though not publicly available) tools for mining the NCI anticancer databases that extends standard COMPARE algorithm.[24] Rabow et al. used Kohonen's self-organizing maps to identify the relationships between chemotypes of screened agents and their effect on four major classes of cellular activities: mitosis, nucleic acid synthesis, membrane transport and integrity, and phosphatase- and kinase-mediated cell cycle regulation.[25] Bykov et al. compared cancer cells growth inhibition profiles for PRIMA-1 (low-molecular weight with antitumor activity) with known anticancer agents and level of p53 protein expression in cancer cells by means of cluster analysis and linear regression.[26] Blower et al. used NCI data in order to find correlations between molecular substructure classes and gene expression data.[27] Marx et al. highlighted the dependence of activity pattern on the structure of *p*-quinones using data mining techniques.[28] Wang and co-workers studied distributions of NCI compounds features in contrast with other chemical databases and reported SMART fragments that are present in active compounds but absent in non-active.[29] Glover et al. revealed new mitochondrial complex I inhibitors with data mining of NCI's anticancer screening database.[30] The NCI data also have been used to predict not only activity, but generalized cytotoxicity as well.[31] The strength of associations of anticancer patterns and levels of molecular expression in these works were measured mainly by Pearson's correlation coefficient. The advantage of such statistics is its unsensitivity to mean activity values, since the same mechanisms can be acted at different levels of effect. But there are also two drawbacks caused mostly by the specific design of both biological and computational experiments. Pearson's correlation coefficient treats all deviations between corresponding elements (cell lines) equally important, while experimental accuracy differs among cell lines. In other words, when performing repeated testings of the same compound, each cell line has its own standard deviation of results sample. To overcome this point we propose to use weighted Pearson's correlation coefficient, adjusting each cell line with the corresponing weight based on standard deviation. The second drawback is the high probability of false findings. Few years ago a broad discussion of this problem took place.[32-35] Indeed, having tens and even hundreds of thousands of correlation coefficients it is easy to find strong correlations without any underlying relationship. And uncertainty in the activity measure (biological assay error) improves the chance. To handle the last point for NCI data, Reinhold et al.[7] proposed to give much confidence to those results where average correlation between repeated measurements is high (more than 0.6), to remove individual probes that has low correlation with other in repeated testing and to consider results from only one experiment less reliable. To decrease the overall likelyhood of false positives we offer to use both high threshold for correlation coefficient values and visualization control. The main argument that makes a claim reliable is replication of that claim. Since the same activity pattern belongs to different (though structurally similar) compounds, their average activity fingerprint is much more close to the real activity pattern and each activity fingerprint of those compounds can be treated as a replicated measure of the same quantity. In other words, the correlation should be treated as significant if it is both high enough and is observed for the whole bunch of similar activity fingerprints. If compared variables represents the same essence (activity fingerprints of compounds in the current case), in order to find significant correlation it is rational to project them on a single plane with dimensionality reduction techniques. And when compared variables have different nature (e.g. activity fingerprint and molecular expression profile), the special kind of correlation plots called "heat map" can be utilized for this purpose. Taking these suggestions into account, we present new protocol for mining of NCI data. Its application on the analysis of 4-thiazolidinones anticancer activity data should bring us closer to the understanding of basic nature of 4-thiazolidinones effect on cancer cells, so that is the main purpose of current artic

2 Data Mining Protocol

The DTP 60-cell line screening is a two-stage process, with the first evaluation of all compounds against the 60 cell lines at a single dose of 10 μM and the second evaluation of only compounds possessing biological activities that exceed a certain threshold value at five doses (including 10 μM). Both assays data have been downloaded from public web source http://dtp.nci.nih.gov/docs/cancer/cancer_data.html (September 2012 release). The results for each compound are reported as the percent growth of treated cells when compared to untreated control cells. It ranges from -100% to 100% (the result of 100% means that treated cells growth is the same as untreated, 0% means that cancer cells stopped growing and -100% means that all treated cancer cells have died. There are concepts of "activity pattern" and "activity fingerprint" often used interchangeably and related to the biological response profile. The next convention is proposed and further utilized in the article. The variation in growth percents of 60 cell lines treated with the same single compound represents activity fingerprint of this compound. Mathematically, it is the set of normalized growth percents (z-scores). So activity fingerprint is the observed property of any compound. Consequently, the set of close activity fingerprints that represents some common activity mechanism is called activity pattern. Activity pattern is the real feature of activity mechanism, but all our knowledge about it comes through activity fingerprints. In other words, we can say about activity pattern when two demands are met: i) similar activity fingerprints are observed for a number of compounds, ii) the overall biological response level of these compounds significantly differs from control testing.

So our major goal is to find anticancer activity patterns among 4-thiazolidinone derivatives compared with other compounds with known activity and also to reveal the relationships between these patterns and cancer cell molecular expression data.

In this study two assumptions are made: i) compounds having the same activity pattern share the same mechanism and ii) the variance of cancer cells growth percents across cell lines exceeds the error of experiment. Both assumptions are supported by a number of publications [4,23,25,26,30].

The proposed protocol of NCI data mining consists of four steps:

1. To estimate the border between active (that can possess some activity pattern) and non-active compounds (no pattern exists);
2. To find the associations between 4-thiazolidinones and NCI public compounds from single-dose dataset and to reveal activity patterns;
3. To find the associations between 4-thiazolidinones and standard anticancer agents;
4. To identify possible molecular targets through comparison between 4-thiazolidinones activity patterns and levels of cancer cell molecular expression.

The methods used at each step are disclosed further in correspondent subsections.

All computations are performed using R 2.15.1[36] – the language and environment for statistical data analysis. For some prespecified tasks additional packages "flexmix" (estimation of Kullback-Leibler divergence),[37] "ggplot2" (figures preparation),[38] "tsne" (dimensionaliy reduction with t-Stochastic Neighbor Embedding),[39] "plyr" (routine data manipulation)[40] and "gplots" (heatmap visualization of relationships between cancer cells gene expression characteristics and activity fingerprints of 4-thiazolidinones)[41] have been used.

2.1 Estimation of the Border between Active and Non-active Compounds

Boyd and Paull argued that the concept of "activity" should be employed only in individualized context for purposes of a given study that contains a decision point(s) dependent upon the assigned definitions.[4] Since the ultimate goal of the majority of papers in anticancer drug development is to discover highly-potent compounds, the "activity" threshold usually is based on the theoretical possibility of medicinal application. Thus the threshold requires that a compound should possess activity at the micromolar level (like in the paper of Fang et al.[24] for example). But in the current study we are interested not in the overall activity, but in the pattern of it. And there are a lot of compounds that are acting through some mechanisms but their activity level is rather low to have some possible practical application. It is rational to involve them into computational analysis in order to increase data amount. To do this we should provide more tolerant threshold. Making an assumption that mean growth percents of cancer cells treated by non-active compound should be normally distributed with center location at 100% (that value represents the same growth of treated cells as untreated are showing), it is possible to find such subset of data that possess empirical distribution closest to theoretical one. This subset will represent non-active compounds, and all others should be treated as active (in the context of current study). For similar purpose earlier we have used Student's t-criterion to test the null hypothesis that subset mean is equal to 100%.[42] This approach solves the problem, but does not take into account the differences in the shapes of empirical and theoretical distributions. Here we propose the more reliable method based on Kullback-Leibler divergence. Particularly, in the case of single dose assay, the Kullback-Leibler divergence between empirical and theoretical distributions is evaluated as function of activity threshold at a grid of 1000 values covering the interval from 80 to 90 percents of cancer cell growth. The desired threshold corresponds to the minimum of this function. This minimum is found to be equal to 0.0993 (some difference in distributions shapes is still present) and is observed at mean growth percent ($\overline{GP}$) = 86.76 that is pretty close to the 86% - the threshold found earlier[42] (the lower precision of later is due to the much smaller dataset). For sure, there are compounds in non-active subset that possess some low anticancer activity and non-active compounds are present in active subset as well. Given threshold should be treated as optimal one that distinguishes two subsets most accurately. Additionally, we had an idea to make the partitioning using standard deviations. Actually, non-active and non-selective cytotoxic compounds should have small standard deviation of testing results (no activity pattern), while active compounds should vary in their activity level across different cell lines. The similar approach was utilized by Rabow et al. in 2002.[25] But the plot of mean activity values *vs* standard deviations (Fig. 1) shows that the difference between standard deviations of highly potent and non-active compounds is rather vague. This plot shows also interesting detail concerning compounds with extremely high cancer cells growth percent. Though two compounds with both high mean growth percent and high standard deviation may represent assay failures (top right corner in the figure), another four compounds located inside the circle (NSC750 – busulfane, NSC178265 – Rhamnolipid R1, NSC303861 - S-(N-methylcarbamate) cysteine ethyl ester monohydrochloride, NSC360036 - neolignan from *Clerodendron inerme*) seems to be cancer cells growth enhancers. Two of the mentioned results are confusing: busulfane is well-known alkylating antineoplastic agent and S-(N-methylcarbamate) cysteine ethyl ester monohydrochloride was reported as possessing some anticancer activity by Jayaram et al.[43]

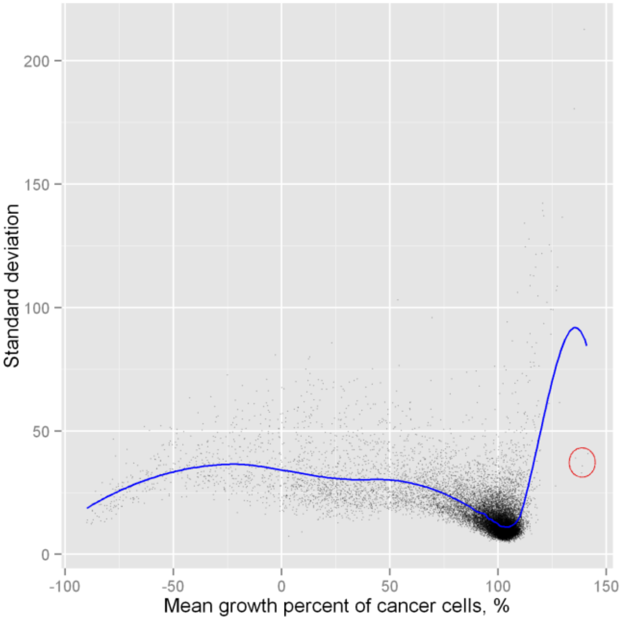


Figure 1. Spatial distribution of mean values and standard deviations from one-dose NCI anticancer assay. Possible cancer cells growth enhancers are outlined with circle.

2.2 Searching for Associations between 4-Thiazolidinones and NCI Public Compounds from Single-Dose Dataset.

3813 Compounds from NCI one-dose assay dataset and 134 compounds from our in-home library of 4-thiazolidinones were considered as active using the found threshold $\overline{GP}$ =86.76%. Pearson's correlation coefficient is the usual statistics to describe the similarity between two activity fingerprints.[6] When using raw growth percent values, Person's correlation coefficient is preferred over other similarity metrics due to its capability to allow (to not take into account) differences in mean activity values. Indeed, Pearson's correlation coefficient is closely related to the squared Euclidean distance between two normalized vectors:

$$D^2(x_{norm}, y_{norm}) = 2n(1-cor(x,y)) \text{ ,}$$

where D^2 is squared Euclidean distance, *n – the length of vectors and cor(x,y) denotes Pearson's correlation coefficient between two raw vectors.*

So the similarity between two activity fingerprints can be estimated using the correlation coefficient between raw growth percents. But this approach has a small drawback: all cell lines are treated equally, while each cell line differs from the others in its sensitivity to chemotherapeutics. In other words, some cell lines show low variability in repeated testings, while the others show high. Thus it is rational to assign weights for each line and calculate weighted Pearson's correlation coefficient. This weights are based on standard deviations (Fig. 2.), obtained from multiple testing of single compound - methotrexate (NSC 740) at 10^{-5} M. The weights are calculated as follows:

$$w_i = \frac{1/\sigma_i}{\sum\limits_i 1/\sigma_i} \text{ ,}$$

where σ_i is the standard deviation of i^{th} cell line.

The standard deviations obtained from multiple testing of single compounds are asymptotically equivalent to the errors of experiment. Additionally, the mean standard deviation has been calculated as square root of mean of cell lines variances (since averaging standard deviations has no sense) and is equal to 14.98. This number represents the standard error of experiment. Comparing the 14.98 with standard deviations of cancer cells growth percents across cell lines (Fig. 1.), we can find only few compounds that show inhibitory effect (are active) and, simultaneously, have standard deviation comparable with standard error of experiment. Therefore, the assumption that the variance of cancer cells growth percents across cell lines exceeds the error of experiment is met for almost all tested compounds.

Returning to the data mining protocol, the pairwise weighted correlation coefficients were obtained for each compounds pair, where the first compound is taken from the NCI one-dose assay dataset and the second – from 4-thiazolidinones library. The distribution of correlation coefficients (Fig. 3.) is almost gaussian with positive shift of center. This may be explained by the presence of the significant amount of non-specific cytotoxics among tested compounds. These cytotoxics together with chance correlations should follow normal distribution. In order to find significant correlations, the minimization of Kullback-Leibler divergence between empirical and theoretical distributions has been utilized once more. This divergence was evaluated as a function of correlation coefficient threshold at a grid of 100 values covering the interval from 0.8 to 0.6. The minimum of divergence is found to be equal to 0.00136 (the empirical distribution is pretty close to gaussian) and is observed at correlation coefficient threshold = 0.700. Again, this threshold does not separate significant correlations from non-significant precisely, but is rather the optimal trade-off between false positives and false negatives.

Then we have filtered the data, leaving only those NCI compounds that correlate at least to one of the 4-thiazolidinones with r>0.700. That results in a correlation matrix of 574 compounds, which was hard to visualize due to overplotting. Moreover, all data visualization techniques were trying to preserve the relations between the compounds with low activity level as well as with high one. Since the number of low-active molecules was much larger, the relationships between highly active compounds were neglected to some extent. These issues became the reason for further data reduction. In this way, only compounds which possess mean growth percent of treated cancer cells < 20 have been retained. The reduced dataset consists of 38 compounds from NCI database and 25 thiazolidinones. The whole data manipulation process is disclosed in Fig. 4.

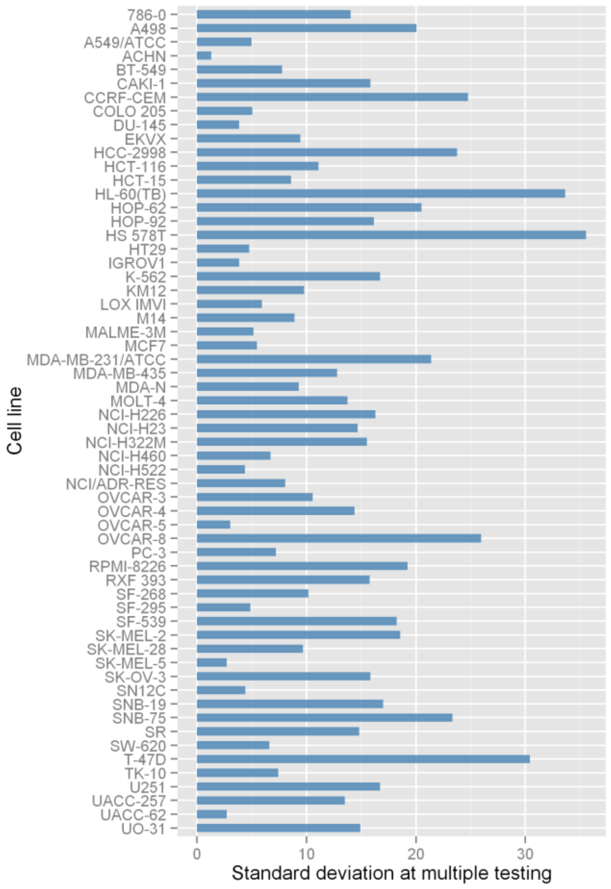


Figure 2. Standard deviations of methotrexate multiple testing results for different cancer cell lines

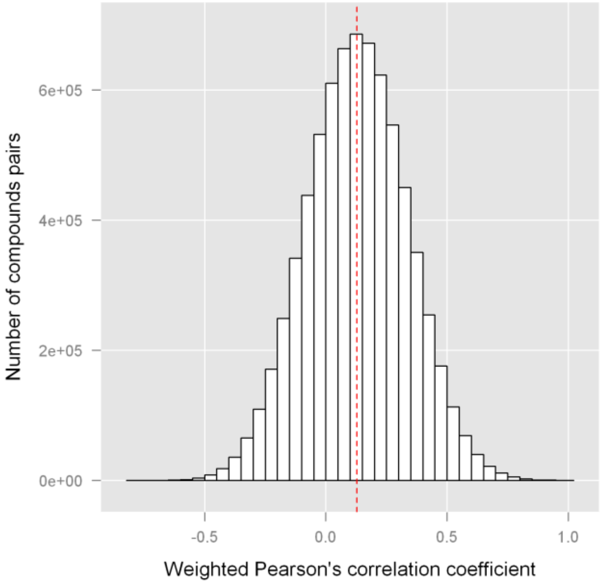


Figure 3. Histogram of correlation coefficients between activity fingerprints of active ($\overline{GP}$ <86.76) compounds from NCI public database and from in-home library of 4-thiazolidinones

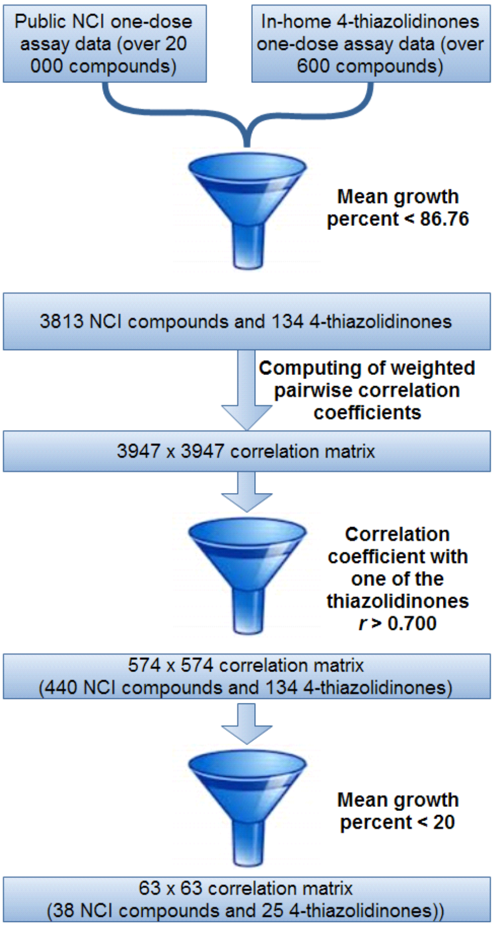

Public NCI one-dose assay data (over 20 000 compounds)
In-home 4-thiazolidinones one-dose assay data (over 600 compounds)
Mean growth percent < 86.76
3813 NCI compounds and 134 4-thiazolidinones
Computing of weighted pairwise correlation coefficients
3947 x 3947 correlation matrix
Correlation coefficient with one of the thiazolidinones r > 0.700
574 x 574 correlation matrix (440 NCI compounds and 134 4-thiazolidinones)
Mean growth percent < 20
63 x 63 correlation matrix (38 NCI compounds and 25 4-thiazolidinones))

Figure 4. The flowchart of data processing and filtering for one-dose assay results mining

It is worth to mention, that visualization of correlation matrix with dimensionality reduction techniques gives the possibility to pinpoint chance correlations, unlike the analysis of raw correlation coefficients does. Specifically, the inferences from such figures are based on joint spatial distribution of data, but not on the single pairwise correlation. Therefore such approach is preferred over the frequentists' statistical tests of significance since it much less suffers from multiple comparison problem.

The rather new method for data visualization called t-SNE (t-Stochastic Neighbor Embedding) has been employed to show the mutual location of activity fingerprints.[39] This technique is based on the representation of distance between datapoints in high-dimensional space into corresponding conditional probabilities of neighborhood between given points and is capable to retain the local structure of the data while also revealing some important global structure (such as clusters at multiple scales). The method proved its efficacy in chemoinformatics, winning the Merck visualization challenge (October 2012) at kaggle.com – the platform for data prediction competitions. It is implemented in R, Python and MATLAB and also can be utilized within CheS-Mapper.[44] The matrix of pairwise correlation distances 1-*r*, where *r* is weighted Pearson correlation coefficient, was the subject of dimensionality reduction with t-SNE, selecting the best result from ten runs (as recommended in the official web-page of the algorithm[45]).

2.3 Searching for Associations between 4-Thiazolidinones and NCI Standard Anticancer Agents

Since compounds from NCI standard agents database are absent in the public results of one-dose anticancer assay, we performed an additional study looking for correlation between 4-thiazolidinones and standard agents activity fingerprints. The activity data for standard agents has been extracted from dose-response database (available at http://dtp.nci.nih.gov/docs/cancer/cancer_data.html) at 10 μM (one-dose assay is carried out with the same concentration) and averaged through several repeated tests. These results were merged with 134 4-thiazolidinones having $\overline{GP}$ < 86.76%. The matrix of pairwise correlation distances have been formed the same way as it is described above and then was submitted to t-SNE visualization algorithm. The flowchart of this process is presented in Fig. 5.

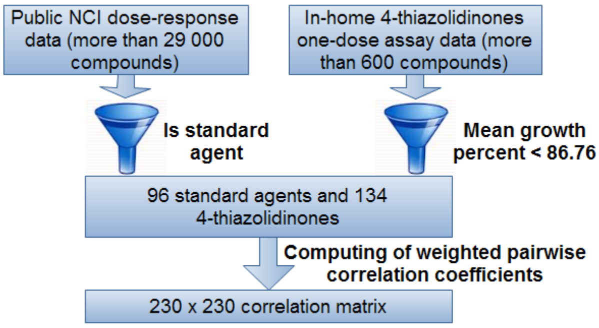

Public NCI dose-response data (more than 29 000 compounds)
In-home 4-thiazolidinones one-dose assay data (more than 600 compounds)
Is standard agent
Mean growth percent < 86.76
96 standard agents and 134 4-thiazolidinones
Computing of weighted pairwise correlation coefficients
230 x 230 correlation matrix

Figure 5. The flowchart of data processing and filtering for standard agents activity mining

2.4 Identification of Possible Molecular Targets for 4-Thiazolidinones with Anticancer Activity

The data used for molecular target search has been downloaded from official NCI web-page.[46] It consists of several datasets:

1. the data from smaller-scale measurements (includes protein, mRNA, miRNA, DNA methylation, mutations, SNPs, enzyme activity, metabolites);
2. cDNA array data from the Weinstein (NCI) and Brown & Botstein (Stanford) groups[23,47];
3. Affymetrix U133 array data from Gene Logic, Inc;
4. Affymetrix U95A data from Novartis, averaged data (from triplicate arrays);
5. Affymetrix U133 array data from Chiron.

In order to estimate the strength of association between molecular characteristics of cancer cells and activity fingerprints the function that describes the relationship should be predefined. This function should be monotonic since the activity can only constantly increase or decrease with the change of molecular characteristics across cell lines but cannot change the direction. For example, it cannot sometimes increase and then sometimes decrease. This is the only preliminary knowledge we can use. Thus, being guided by Occam's razor principle, one of the simplest monotonic functions – the linear function has been chosen as the link between two studied variables. So the weighted Pearson's correlation coefficient has been utilized in this case as well.

The correlation matrices of different characteristics of NCI-60 tumor cell lines versus 134 4-thiazolidinones possessing $\overline{GP}$ < 86.76% have been obtained for each dataset. The characteristics that have no any high correlation coefficient with at least one compound's activity fingerprint were removed, and the resulting matrices were visualized by heat maps.

3. Results and Discussion

3.1 Searching for Associations between 4-Thiazolidinones and NCI Public Compounds from Single-Dose Dataset.

Visualization of activity fingerprints (Fig. 6) shows that there are at least 5 distinct patterns among studied compounds. Since all NCI compounds in the figure were selected as those showing high correlation with at least one of the 4-thiazolidinones, and there is a group of NCI compounds far from any 4-thiazolidinone (located at the left part of Fig. 6), we can conclude that this case represents the issue of chance correlations. In fact, i) several 4-thiazolidinones are correlated with above mentioned NCI compounds (their activity fingerprints are correlated to be more exactly); ii) these NCI compounds are correlated with each other; iii) the same 4-thiazolidinones are strongly correlated with other compounds, while the NCI compounds do not. All this together is forming the picture, where the left and partially the right groups consist of NCI compounds only and are located separately from others. The structures of other NCI compounds were subjected to the literature search and the most interesting findings are identified in the plot (Fig. 6) and listed in Table 1 together with the closest 4-thiazolidinones.

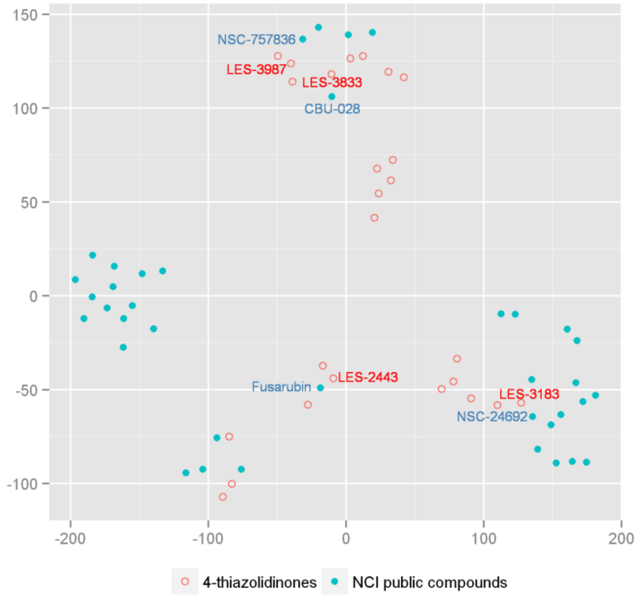
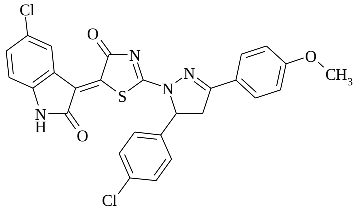
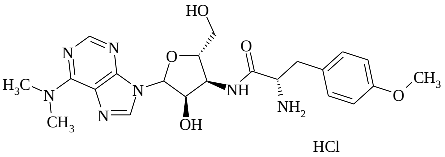
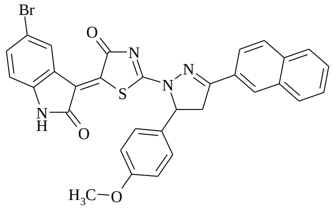
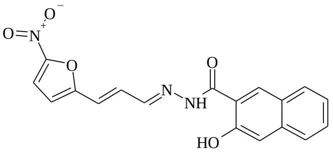
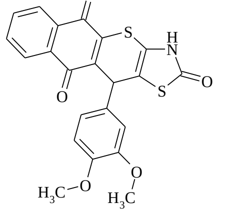
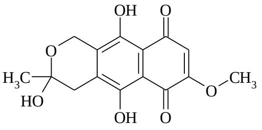


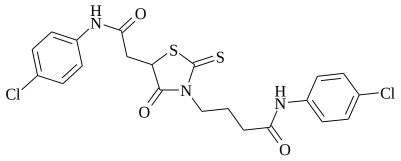
Figure 6. Visualization of activity fingerprints. Compounds from in-home database are represented with hollow circles, while compounds from NCI public one-dose assay data are represented with filled ones.

NSC-757836 is an optical isomer of puromycin that inhibits protein synthesis by binding to RNA and causing premature release of protein from the synthetic site of the ribosome.[48,49] Les-3987 activity fingerprint is highly correlated with NSC-757836, assuming the same mode of antiproliferative action. The activity fingerprint of CBU-028 is located rather close to this pair. CBU-028 is a bifunctional compound with a membrane-binding domain and a p97/VCP inhibiting activity, disrupting protein homeostasis in endoplasmic reticulum. Les-3833 is the closest compound to CBU-028 with weighted correlation coefficient *r* = 0.712. Given results suggest that 4-thiazolidinones conjugated with substituted pyrazoline and isatin fragments act through inhibition of protein synthesis, but this inhibition may be conducted with two different mechanisms. There is also a group of five similar 4-thiazolidinones possessing the same structural template as Les-3987 and Les-3833 and located separately but not so far from CBU-028. In our opinion, this point does not support the hypothesis about some other activity mechanism, but should be treated as algorithm's artifact. Actually, the perplexity parameter of t-SNE was equal to 5. In t-SNE, the perplexity may be viewed as a knob that sets the number of effective nearest neighbors.[45] This way five similar 4-thiazolidinones formed a compact single group weakly interacting with other neighbors. So the activity mechanism of this group can be the same as other 4-thiazolidinones conjugated with substituted pyrazoline and isatin fragments have. Further, the same mode of action is observed for fusarubin and thiopyrano[2,3-d][1,3]thiazol-2-ones with similar naphtoquinone fragment. Fusarubin is the pigment isolated from *Fusarium solani* possessing antibacterial and antitumor properties.[50] This finding encourages to test the antibacterial activity of mentioned thiopyrano[2,3-d][1,3]thiazol-2-ones. The strongest correlation (*r* = 0.792) was observed between Les-3183 and NSC-24692. NSC-24692 is 2,4,7-tribromotropone and it is found to be active in 48 assays from 228 listed in PubChem, while Les-3183 is rhodanine derivative with two peptide bonds. Taking into account complete dissimilarity of these two structures and the fact that rhodanines are known as promiscuous compounds[51] we can assume that both compounds and maybe their neighbors as well are pan assay interference compounds.

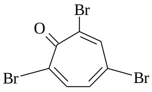
It is worth to notice that among given 38 highly potent NCI public compounds whose activity fingerprints strongly (*r*>0.700) correlates with 4-thiazolidinones fingerprints more than a half are really located far from 4-thiazolidinones on the visualization plane (Fig. 6). Using common analysis of raw correlation coefficients, the relationships with these compounds could be (mistakenly) treated as significant. In such a way the proposed visualization helps us to avoid false discoveries.

Table 1. 4-Thiazolidinones and Compounds from NCI Public Database Showing Similar Activity Fingerprints

4-Thiazolidinone derivatives	NCI public compounds	Weighted correlation coefficient
 Les-3987, $\overline{GP}$=-58.01	 NSC-757836, optical isomer of puromycin – well-known anticancer agent that inhibits protein synthesis by binding to RNA, $\overline{GP}$=-44.28	r=0.741
 Les-3833, $\overline{GP}$=-59.81	 CBU-028, bifunctional compound with a membrane-binding domain and a p97/VCP inhibiting activity,[52] $\overline{GP}$=-36.64	r=0.712
 Les-2443, $\overline{GP}$= 1.76	 Fusarubin is the pigment isolated from *Fusarium solani* possessing antibacterial and antitumor properties, $\overline{GP}$=5.93	r=0.722



Les-3183, $\overline{GP}$ =-7.32



r=0.792

NSC-24692, 2,4,7-Tribromotropone, $\overline{GP}$ =-18.27
(Pubchem: tested in 228 assays, active in 48).

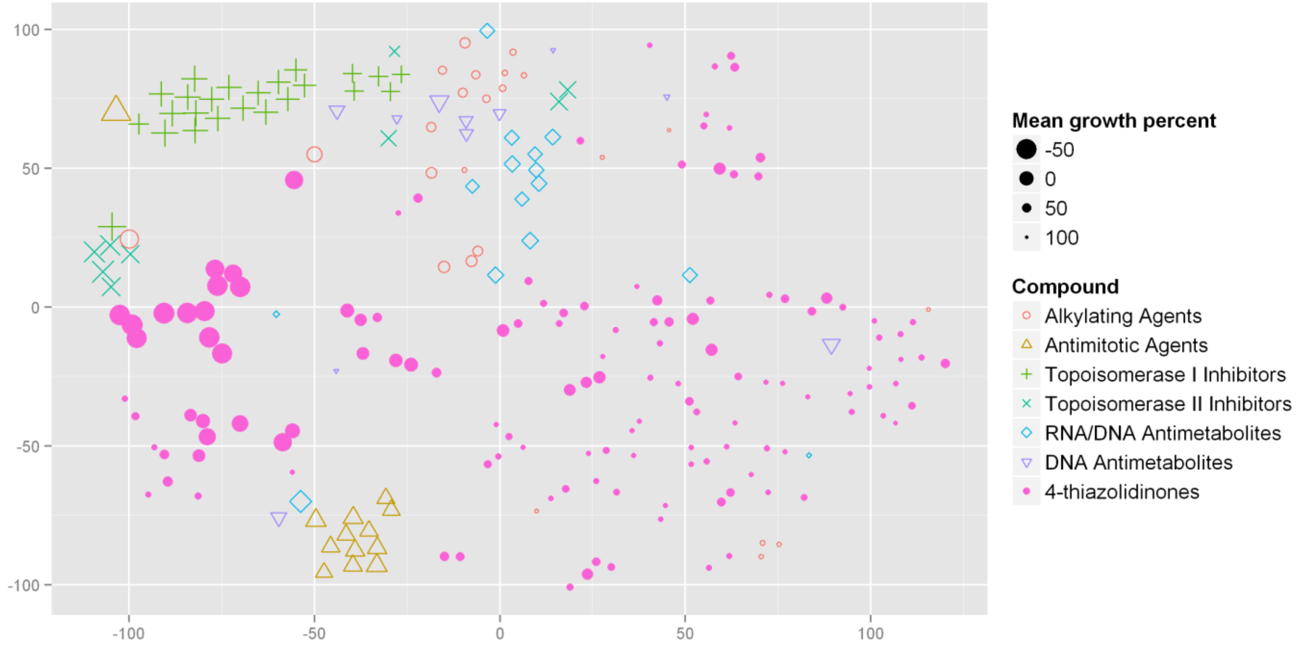


Figure 7. Activity fingerprints of standard agents and 4-thiazolidinones.

3.2 Searching for Associations between 4-Thiazolidinones and NCI Standard Anticancer Agents

The visualization (Fig. 7) fairly distinguishes different mechanisms of standard agents (except few non-trivial compounds). Additionally, some standard agents (especially antimitotics) have noticeably small mean activity values, indicating that average cancer cells growth percent is not a perfect outcome for design of new antitumor medicines. We can see also that low-active 4-thiazolidinones activity templates are diverse and differ from those of standard agents. Moreover, 4-thiazolidinones with high activity show distinct pattern also. Several 4-thiazolidinones (containing pyrazoline and isatin fragments) are close to topoisomerase II inhibitors, but the corresponding correlation coefficients lie within the range 0.51-0.67 and none of them exceeds 0.70. So, 4-thiazolidinones do not show any of the standard anticancer mechanisms.

3.3 Identification of Possible Molecular Targets for 4-Thiazolidinones with Anticancer Activity

Searching for associations between the data from smaller-scale measurements and 4-thiazolidinones activity fingerprints, we have found that the most active group of compounds (the first one from the left in Fig. 8.) is correlated with methylation of CpG sites in CD34 (MT13805), AF5q31 - ALL1 fused gene from chromosome 5q31 (MT9796), and SYK (MT8619) (Fig. 8). This group consists of 15 4-thiazolidinones with pyrazoline and isatin fragments. The methylation data comes from M. Ehrich et al.[53] and represents fraction of DNA methylated. Since low cancer cells growth percents are related to high activity, the high methylated fraction of CD34 and AF5q31 in cancer cells makes the cells more sensible to compounds of interest, but the influence of SYK methylated fraction is reverse. The protein encoded by AF5q31 (also known as AF4/FMR2 family, member 4) belongs to the AF4 family of transcription factors involved in leukemia. The SYK gene encodes spleen tyrosine kinase which is widely expressed in hematopoietic cells and is involved in coupling activated immunoreceptors to downstream signaling events that mediate diverse cellular responses, including proliferation, differentiation, phagocytosis and cell-cell adhesion.[54] Spleen tyrosine kinase can activate the oncogenic transcription factor “The Signal Transducer and Activator of Transcription 3” (STAT3) to induce expression of STAT3 target genes that improves resistance of human B-lineage leukemia/lymphoma cells to oxidative stress-induced apoptosis.[55] The group of 4-thiazolidinones with pyrazoline and isatin fragments includes subgroup (the first 4 compounds in Fig. 9) that shows strong negative correlation with methylated fraction of CpG sites in HOXA5 (MT4654). Hypermethylation of HOXA5 lowers its expression and is observed in clear cell renal cell carcinoma,[56] acute myeloid leukemia,[57] oral squamous cell carcinoma,[58] non-small cell lung cancer[59] etc, and in all cases were related to poor prognosis. These 4 compounds are listed in table 2 and were found to be highly active exactly in cells with large fraction of methylated HOXA5.

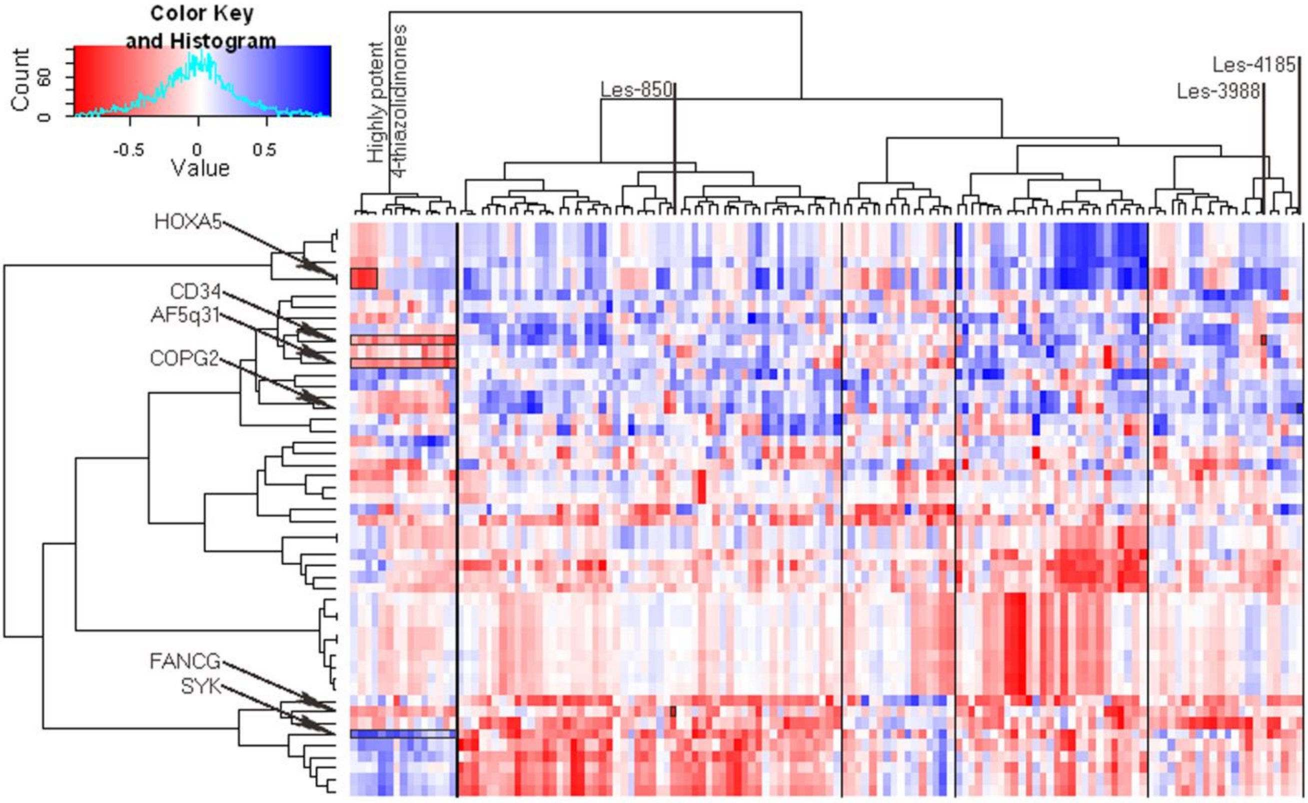
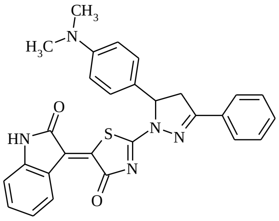
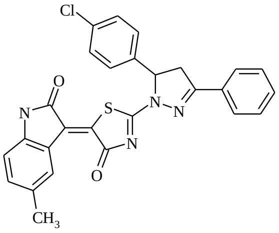
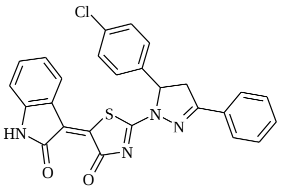
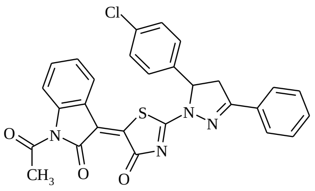


Figure 8. The correlation matrix of 4-thiazolidinones activity fingerprints versus NCI-60 cell characteristics obtained from smaller-scale measurements. The columns correspond to compounds activities while rows represent fractions of methylated CpG sites in different genes.

Table 2. Compounds Showing High Negative Correlation Between Their Activity Fingerprint and Fraction of Methylated CpG Sites in HOXA5.

ID	Structure	$\overline{GP}$	Weighted correlation coefficient
Les-3640		-40	-0.622
Les-3639		-54	-0.763
Les-3643		-57	-0.712
Les-3645		-40	-0.657

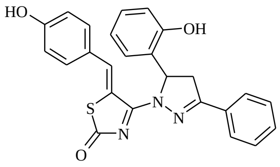
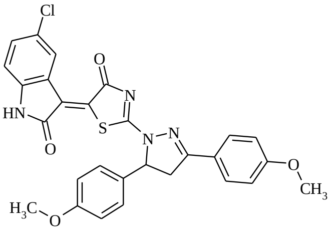
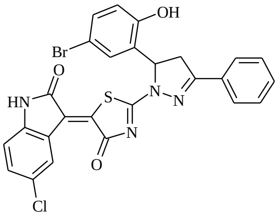
Similar activity fingerprints of Les-3643 and Les-3645 support our previous hypothesis about the hydrolysis of N-acethyl fragment prior to interaction with biotarget.[42]

There are few other single high correlations observed for highly potent 4-thiazolidinones (Table 3.). Remarkably that the thiopyrano[2,3-*d*][1,3]thiazol-2-one and 2-thioxo-4-thiazolidinone (rhodanine) derivatives (Les-2443 and Les-3183 respectively) do not show any significant relationships with studied cancer cells characteristics.

Taking cDNA microarray data,[23,47] much less relationships have been identified (Fig. 9). The most active group of compounds (the first one in Fig. 10) shows common, though not very strong positive correlation with malonyl CoA:ACP acyltransferase (mitochondrial) (GC13049) and IFI27- interferon alpha-inducible protein 27 (GC13627). The last is overexpressed in patients with chronic myeloproliferative neoplasms.[60] The sign of correlation coefficients suggests that high expression of both genes is related to lower activity levels of these compounds. Les-3183 (rhodanine derivative) was found to be correlated (*r*=0.847) with interferon alpha-inducible protein 27 as well.

The heat maps obtained with mRNA expression data from GeneLogic Inc, Novartis and Chiron are almost identical and does not highlight any significant findings. These figures are provided as supporting information.

Table 3. The Strongest Correlations Found Between 4-Thiazolidinones Activity Fingerprints and Cancer Cells Characteristics Obtained from Smaller-Scale Measurements.

ID	Structure	$\overline{GP}$	Cell characteristics	Weighted correlation coefficient
Les-850		1	Fraction of methylated CpG sites in FANCG (MT5858)	-0.816
Les-3988		1	Fraction of methylated CpG sites in CD34 (MT13805)	-0.845
Les-4185		19	Fraction of methylated CpG sites in COPG2 (MT5479)	0.871

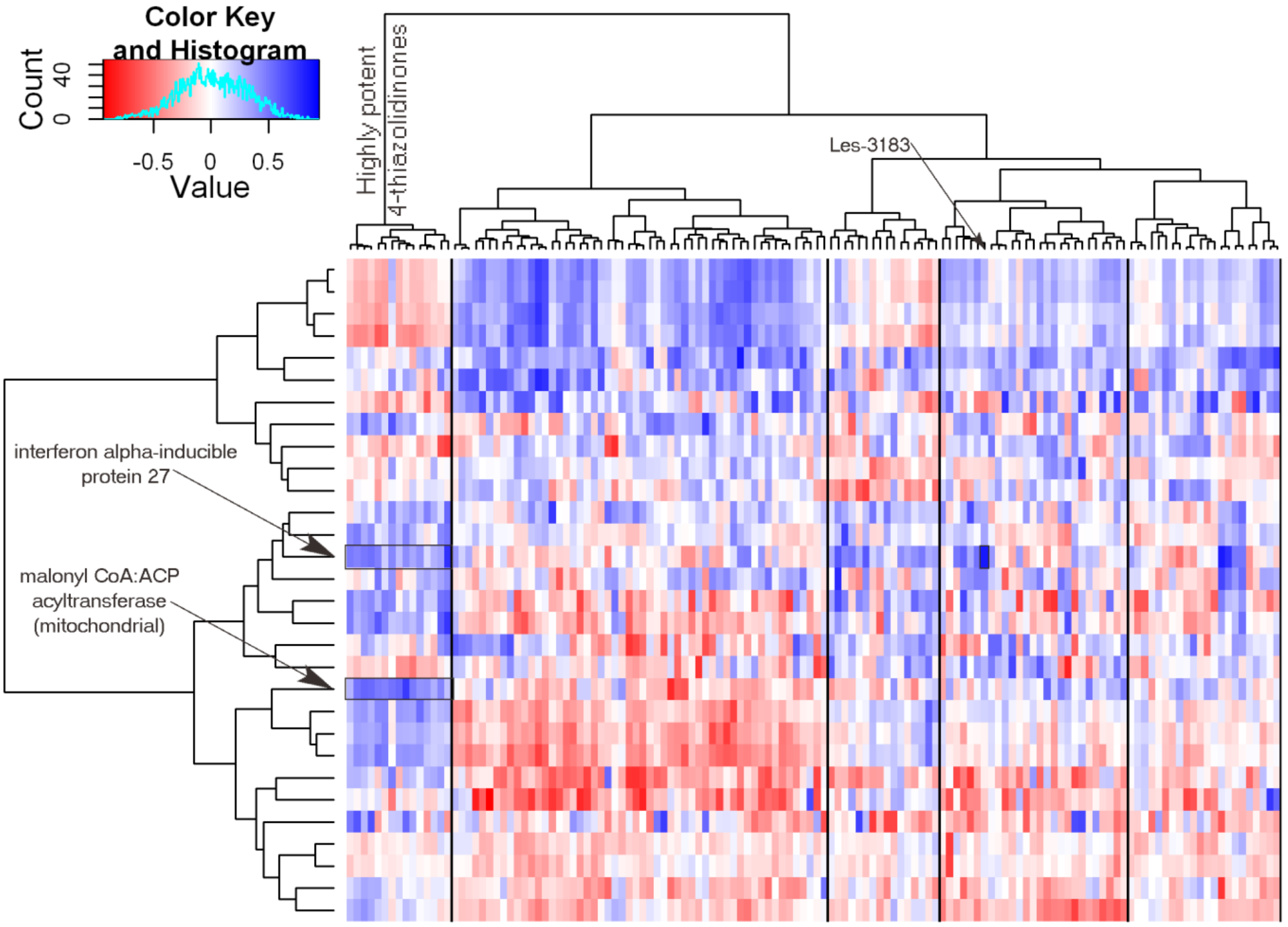


Figure 9. The correlation matrix of 4-thiazolidinones activity fingerprints versus NCI-60 cell characteristics obtained from cDNA microarray. The columns correspond to compounds activities while rows represent log(mRNA levels in cell line/mRNA levels in reference pool).

3 Conclusions

The proposed protocol of NCI data mining shows its capability to find significant associations and relationships while filtering out high correlations produced by chance. The visualization of results is a powerful approach for decision making. Using t-SNE algorithm, the group of highly potent 4-thiazolidinone-pyrazoline-isatin conjugates was found to have common pattern of antitumor activity. This pattern is significantly close to puromycin and a novel promising anticancer agent CBU-028. Both of them inhibit protein synthesis, but in different ways: the pyromucin disrupts protein formation while CBU-028 – protein degradation, halting protein synthesis through negative feedback loop. The strong correlation of these compounds activity fingerprints with fraction of methylated CpG sites in CD34, AF5q31 and SYK have been observed. AF5q31 expression plays important role in tumor cells proliferation and metastasis, while SYK is novel promising target for the development of anticancer agents. There is also common, though not very strong positive correlation of current group with malonyl CoA:ACP acyltransferase (mitochondrial) and IFI27 - interferon alpha-inducible protein 27. Several compounds from this group show strong negative correlation with methylated fraction of CpG sites in HOXA5, currently viewed as prognostic marker for a set of malignant tumors. These results outline further directions of biochemical studies. Additionally, thiopyrano[2,3-d][1,3]thiazol-2-ones bearing naphtoquinone fragment have been found to possess the same activity pattern as fusarubin does. Fusarubin is an antibiotic with antitumor properties. So the hypothesis about antibacterial properties of such thiopyrano[2,3-*d*][1,3]thiazol-2-one derivatives is useful side result of current computational study. One of the other highly potent 4-thiazolidinone derivatives, namely Les-3183 seems to be pan assay interference compound. Generally, none of the studied 4-thiazolidinones has activity fingerprint similar to standard anticancer agents. The obtained results produce new hypotheses about the routes of 4-thiazolidinones anticancer effect. Confirmation of them will bring medicinal chemistry closer to the development of new 4-thiazolidinone-based anticancer drug.

Supporting information

The correlation matrices of 4-thiazolidinones activity fingerprints versus different cancer cells characteristics visualized with heat maps. These characteristics are the follows:

Fig. S1 - the data from smaller-scale measurements (includes protein, mRNA, miRNA, DNA methylation, mutations, SNPs, enzyme activity, metabolites);
Fig. S2 - cDNA array data from the Weinstein (NCI) and Brown & Botstein (Stanford) groups;
Fig. S3 - Affymetrix U133 array data from Gene Logic, Inc;
Fig. S4 - Affymetrix U95A data from Novartis, averaged data (from triplicate arrays);
Fig. S5 - Affymetrix U133 array data from Chiron.

Translation of NCI target pattern identifiers (row names) into corresponding gene information can be performed via NCI web-service http://dtp.nci.nih.gov/mtweb/search.jsp

References

[1] G. M. B. Cragg, M.R.; Khanna, R.; Kneller, R; Mays T.D.; Mazan, K.D.; Newman, D.J.; Sausville E.A., *Pure Appl. Chem.* **1999,** 71, 1619-1633.

[2] R. H. Shoemaker, *Nat. Rev. Cancer* **2006,** 6, 813-823.

[3] Compare Main Page. http://dtp.nci.nih.gov/compare/ (22.12.2012)

[4] M. R. Boyd, K. D. Paull, *Drug Dev. Res.* **1995,** 34, 91-109.

[5] K. D. Paull, R. H. Shoemaker, L. Hodes, A. Monks, D. A. Scudiero, L. Rubinstein, J. Plowman, M. R. Boyd, *J. Natl. Cancer Inst.* **1989,** 81, 1088-1092.

[6] D. W. Zaharevitz, S. L. Holbeck, C. Bowerman, P. A. Svetlik, *J. Mol. Graph. Model.* **2002,** 20, 297-303.

[7] W. C. Reinhold, M. Sunshine, H. Liu, S. Varma, K. W. Kohn, J. Morris, J. Doroshow, Y. Pommier, *Cancer Res.* **2012,** 72, 3499-3511.

[8] G. Marzaro, A. Chilin, A. Guiotto, E. Uriarte, P. Brun, I. Castagliuolo, F. Tonus, H. González-Díaz, *Eur. J. Med. Chem.* **2011,** 46, 2185-2192.

[9] A. Speck-Planche, V. V. Kleandrova, F. Luan, M. Natalia D. S. Cordeiro, *Anti-Cancer Agents in Medicinal Chemistry* **2012,** 12, 678.

[10] A. Speck-Planche, V. V. Kleandrova, F. Luan, M. N. D. S. Cordeiro, *Anti-Cancer Agents in Medicinal Chemistry* **2013,** 13, 791-800.

[11] D. Havrylyuk, B. Zimenkovsky, R. Lesyk, *Phosphorus, Sulfur Silicon Relat. Elem.* **2009,** 184, 638-650.

[12] R. Lesyk, B. Zimenkovsky, D. Kaminskyy, A. Kryshchyshyn, D. Y. Havryluk, D. Atamanyuk, I. Y. Subtel'na, D. Khyluk, *Biopolymers and Cell* **2011,** 27, 107-117.

[13] R. Lesyk, B. Zimenkovsky, D. Atamanyuk, F. Jensen, K. Kiec-Kononowicz, A. Gzella, *Biorg. Med. Chem.* **2006,** 14, 5230-40.

[14] A. Kryshchyshyn, D. Atamanyuk, R. Lesyk, *Sci Pharm* **2012,** 80, 509-29.

[15] D. Havrylyuk, B. Zimenkovsky, O. Vasylenko, L. Zaprutko, A. Gzella, R. Lesyk, *Eur. J. Med. Chem.* **2009,** 44, 1396-404.

[16] I. Subtel'na, D. Atamanyuk, E. Szymanska, K. Kiec-Kononowicz, B. Zimenkovsky, O. Vasylenko, A. Gzella, R. Lesyk, *Biorg. Med. Chem.* **2010,** 18, 5090-102.

[17] D. Havrylyuk, B. Zimenkovsky, O. Vasylenko, A. Gzella, R. Lesyk, *J. Med. Chem.* **2012,** 55, 8630-41.

[18] D. Havrylyuk, L. Mosula, B. Zimenkovsky, O. Vasylenko, A. Gzella, R. Lesyk, *Eur. J. Med. Chem.* **2010,** 45, 5012-21.

[19] D. Kaminskyy, D. Khyluk, O. Vasylenko, L. Zaprutko, R. Lesyk, *Sci Pharm* **2011,** 79, 763-77.

[20] L. M. Shi, T. G. Myers, Y. Fan, P. M. O'Connor, K. D. Paull, S. H. Friend, J. N. Weinstein, *Mol. Pharmacol.* **1998,** 53, 241-251.

[21] L. M. Shi, Y. Fan, T. G. Myers, P. M. O'Connor, K. D. Paull, S. H. Friend, J. N. Weinstein, *J. Chem. Inf. Comput. Sci.* **1998,** 38, 189-199.

[22] L. M. Shi, Y. Fan, J. K. Lee, M. Waltham, D. T. Andrews, U. Scherf, K. D. Paull, J. N. Weinstein, *J. Chem. Inf. Comput. Sci.* **1999,** 40, 367-379.

[23] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, J. N. Weinstein, *Nat. Genet.* **2000,** 24, 236-244.

[24] X. Fang, L. Shao, H. Zhang, S. Wang, *J. Chem. Inf. Comput. Sci.* **2003,** 44, 249-257.

[25] A. A. Rabow, R. H. Shoemaker, E. A. Sausville, D. G. Covell, *J. Med. Chem.* **2002,** 45, 818-840.

[26] V. J. N. Bykov, N. Issaeva, G. Selivanova, K. G. Wiman, *Carcinogenesis* **2002,** 23, 2011-2018.

[27] P. E. Blower, C. Yang, M. A. Fligner, J. S. Verducci, L. Yu, S. Richman, J. N. Weinstein, *Pharmacogenomics J.* **2002,** 2, 259-271.

[28] K. A. Marx, P. O'Neil, P. Hoffman, M. L. Ujwal, *J. Chem. Inf. Comput. Sci.* **2003,** 43, 1652-1667.

[29] H. Wang, J. Klinginsmith, X. Dong, A. C. Lee, R. Guha, Y. Wu, G. M. Crippen, D. J. Wild, *J. Chem. Inf. Model.* **2007,** 47, 2063-2076.

[30] C. J. Glover, A. A. Rabow, Y. G. Isgor, R. H. Shoemaker, D. G. Covell, *Biochem. Pharmacol.* **2007,** 73, 331-340.

[31] A. C. Lee, K. Shedden, G. R. Rosania, G. M. Crippen, *J. Chem. Inf. Model.* **2008,** 48, 1379-1388.

[32] J. P. A. Ioannidis, *Trends in Molecular Medicine* **2003,** 9, 135-138.

[33] J. P. A. Ioannidis, *PLoS Med* **2005,** 2, e124.

[34] R. Moonesinghe, M. J. Khoury, A. C. J. W. Janssens, *PLoS Med* **2007,** 4, e28.

[35] B. Djulbegovic, I. Hozo, *PLoS Med* **2007,** 4, e26.

[36] R. Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.

[37] B. Grün, F. Leisch, *J. Stat. Softw.* **2008,** 28, 1-35.

[38] H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis. 2009*. Springer: New York.

[39] L. Van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008,** 9, 2579-2605.

[40] H. Wickham, *J. Stat. Softw.* **2011,** 40, 1–29.

[41] G. R. Warnes Gplots: Various R Programming Tools for Plotting Data. http://CRAN.R-project.org/package=gplots (12.01.2013)

[42] O. Devinyak, B. Zimekovsky, R. Lesyk, *Curr. Top. Med. Chem.* **2012,** 12, 2763-2784.

[43] H. N. Jayaram, M. S. Lui, J. Plowman, K. Pillwein, M. A. Reardon, W. L. Elliott, G. Weber, *Cancer Chemother. Pharmacol.* **1990,** 26, 88-92.

[44] M. Gütlein, A. Karwath, S. Kramer, *J. Cheminf.* **2012,** 4, 1-16.

[45] L. J. P. van_der_Maaten The Official Web-Page of t-SNE. http://homepage.tudelft.nl/19j49/t-SNE.html (29.12.2012)

[46] DTP - Download Page for Molecular Target Data. http://dtp.nci.nih.gov/mtargets/download.html (10.01.2013)

[47] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, *Nat. Genet.* **2000,** 24, 227-235.

[48] M. Rabinovitz, J. M. Fisher, *J. Biol. Chem.* **1962,** 237, 477-481.

[49] D. Nathans, *Proc. Natl. Acad. Sci. U. S. A.* **1964,** 51, 585.

[50] H. J. Issaq, E. W. Barr, T. Wei, C. Meyers, A. Aszalos, *J. Chromatogr. A* **1977,** 133, 291-301.

[51] T. Tomašić, L. P. Mašič, *Expert Opin. Drug Discovery* **2012,** 7, 549-560.

[52] Q. Wang, B. A. Shinkre, J.-g. Lee, M. A. Weniger, Y. Liu, W. Chen, A. Wiestner, W. C. Trenkle, Y. Ye, *PloS one* **2010,** 5, e15479.

[53] M. Ehrich, J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, D. van den Boom, *Proc. Natl. Acad. Sci. U. S. A.* **2008,** 105, 4844-9.

[54] R. Larive, S. Urbach, J. Poncet, P. Jouin, G. Mascré, A. Sahuquet, P. Mangeat, P. Coopman, N. Bettache, *Oncogene* **2009,** 28, 2337-2347.

[55] F. M. Uckun, S. Qazi, H. Ma, L. Tuel-Ahlgren, Z. Ozer, *Proc. Natl. Acad. Sci. U. S. A.* **2010,** 107, 2902-2907.

[56] K. H. Yoo, Y. K. Park, H. S. Kim, W. W. Jung, S. G. Chang, *Pathol. Int.* **2010,** 60, 661-666.

[57] S. Y. Kim, S. H. Hwang, E. J. Song, H. J. Shin, J. S. Jung, E. Y. Lee, *Korean J. Lab. Med.* **2010,** 30, 469-473.

[58] C. O. Rodini, F. C. A. Xavier, K. B. S. Paiva, M. De Souza Setóbal Destro, R. A. Moyses, P. Michaluarte, M. B. Carvalho, E. E. Fukuyama, E. H. Tajara, O. K. Okamoto, F. D. Nunes, *Int. J. Oncol.* **2012,** 40, 1180-1188.

[59] D. S. Kim, M. J. Kim, J. Y. Lee, S. M. Lee, J. Y. Choi, G. S. Yoon, Y. K. Na, H. S. Hong, S. G. Kim, J. E. Choi, *Mol. Carcinog.* **2009,** 48, 1109-1115.

[60] V. Skov, T. S. Larsen, M. Thomassen, C. H. Riley, M. K. Jensen, O. W. Bjerrum, T. A. Kruse, H. C. Hasselbalch, *Eur. J. Haematol.* **2011,** 87, 54-60.

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))