Published online 8 October 2008

Nucleic Acids Research, 2008, Vol. 36, No. 20 ***e137***
doi:10.1093/nar/gkn610

Gene expression module-based chemical function similarity search

Yun Li[1,2], Pei Hao[1,2], Siyuan Zheng[1], Kang Tu[1], Haiwei Fan[2], Ruixin Zhu[2], Guohui Ding[1], Changzheng Dong[1], Chuan Wang[1], Xuan Li[2], H.-J. Thiesen[3], Y. Eugene Chen[4], Hualiang Jiang[5], Lei Liu[1] and Yixue Li[1,2,*]

[1]Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, [2]Shanghai Center for Bioinformation Technology, F1.12, No.100, Qinzhou Road, Shanghai 200235, PR China, [3]Institute of Immunology of Rostock University, Schillingallee 69, D-18055 Rostock, Germany, [4]Cardiovascular Center, Department of Internal Medicine, University of Michigan Medical Center, Ann Arbor, MI 48109, USA and [5]Drug Discovery and Design Centre, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, PR China

Received May 20, 2008; Revised September 8, 2008; Accepted September 9, 2008

ABSTRACT

Investigation of biological processes using selective chemical interventions is generally applied in biomedical research and drug discovery. Many studies of this kind make use of gene expression experiments to explore cellular responses to chemical interventions. Recently, some research groups constructed libraries of chemical related expression profiles, and introduced similarity comparison into chemical induced transcriptome analysis. Resembling sequence similarity alignment, expression pattern comparison among chemical intervention related expression profiles provides a new way for chemical function prediction and chemical–gene relation investigation. However, existing methods place more emphasis on comparing profile patterns globally, which ignore noises and marginal effects. At the same time, though the whole information of expression profiles has been used, it is difficult to uncover the underlying mechanisms that lead to the functional similarity between two molecules. Here a new approach is presented to perform biological effects similarity comparison within small biologically meaningful gene categories. Regarding gene categories as units, a reduced similarity matrix is generated for measuring the biological distances between query and profiles in library and pointing out in which modules do chemical pairs resemble. Through the modularization of expression patterns, this method reduces experimental noises and marginal effects and directly correlates chemical molecules with gene function modules.

INTRODUCTION

Exploring the cellular responses to chemicals is practically meaningful in biomedical research and drug discovery. Microarray technology, due to its potential for monitoring genome-wide expression changes in response to chemical interventions, has applied to many endeavors in chemical biology research, including chemical toxicity investigation (1,2), chemical target discovery (3,4) and chemical regulated pathway identification (5).

Meanwhile some efforts have also been made to construct large-scale libraries of expression profiles corresponding to diverse chemical treatments. Hughes *et al.* (6) produced a library of expression profiles corresponding to diverse mutations and chemical treatments in *Saccharomyces cerevisiae*. They illustrated for the first time the utility of transcriptome data in identification as well as functional classification of unknown genes. In Fielden *et al.* (7) and Nie *et al.*'s (8) work, combining with results from 2-year rodent bioassay, microarray data of chemical treated rats was used to select gene biomarkers that distinguish carcinogenic chemicals from noncarcinogenic ones. Lamb *et al.* (9), on the other hand, established a searchable database of expression profiles corresponding to human cell lines treated with diverse

*To whom correspondence should be addressed. Tel: +86 21 54065060; Fax: +86 21 54065058; Email: yxli@sibs.ac.cn
Correspondence may also be addressed to Lei Liu. Email: liulei@scbit.org; Hualiang Jiang. Email: hljiang@mail.shcnc.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

chemical molecules. They devised a method for global expression pattern comparison, and constructed a searchable system, Connectivity Map, which survey the identities of biological effects of chemicals by performing profile similarity search. Ideas and systems related to Lamb's work have been used by many research groups and proved helpful in potential therapeutic agent discovery (10) and in identifying pathways regulated by small molecules (11).

Lamb *et al*. unprecedentedly introduced the concept of 'similarity search' into chemical related transcriptome analysis and emphasized global pattern comparison of gene expression profiles. Based on the holistic information of gene expression profiles, Lamb's method provided a new way to decipher the functional relationship among small chemical molecules even if they have different structures. However, it is not possible for researchers to see more deeply into the underlying biology, i.e. in which biological processes the given chemicals are involved in, and how do they result in analogs regulatory mechanisms. If we can shed light on these problems, we may have a chance to understand how the human body handles drugs and how side effects of a drug take place.

Based on those considerations and starting from Connectivity Map (9), we developed a new approach to perform functional similarity search for chemical molecules. To be different from global expression pattern comparison Lamb *et al*. used, we emphasized on comparing expression patterns in each gene module. The concept of 'gene module' here refers to a set of genes that act in concert to carry out a specific function (12). For example, a group of genes involves in cell cycle can be defined as a gene module that participates in cell cycle regulation. So far, many rules have been set from different perspectives to compile genes into biologically meaningful categories, like pathway information (13,14) or function annotations. Gene ontology (GO) (15), with its effort on developing structured vocabularies in describing and classifying gene products, is widely used in exploring biological features of genes with respect to molecular functions, biological processes as well as cellular components. It has also been proved useful in dealing with microarray data, including providing functional annotation of genes observed differentially expressed, gaining insight into the underlying biological mechanisms (16,17) and grouping microarray data according to the functions of the genes or biological processes they are involved in Ref. (18). Considering its advantages, we chose GO as the rule to group genes into units. Each genes unit is called a gene ontology module (GOM) representing a group of functionally associated genes. Instead of taking all genes' expression pattern into account, we restricted our expression pattern comparison into every GOM. Regarding GOMs as units, a reduced similarity matrix was generated for measuring the biological distances between query profile and profiles in library and pointing out in which modules the chemical pairs resemble. Like Coarse-Graining Approaches (CGA) (19) in theoretical physics, our strategy smoothes over fine detail and extracts crucial elements from overwhelming information. Through the modularization of expression patterns, this method reduces experimental noises and marginal effects and directly correlates small chemical molecules with gene function modules. In our article some cases have been tested to show that our method is sensitive, and can provide reasonable results.

MATERIALS AND METHODS

Data source

Data in reference library was downloaded from Connectivity Map (build 01) (http://www.broad.mit.edu/cmap_build01/). It consists of 564 gene expression profiles corresponding to human cultured cell lines treated with 164 distinct chemical molecules representing a total of 453 instances. Each instance here denotes a treatment and vehicle pair.

Data preprocessing

Raw data were first normalized [RMA (20)] and log transformed. Each instance was then processed using the following three steps:

Step 1: Log ratio of treatment to vehicle (mean) was calculated for each probe;
Step 2: All probes were then mapped to the corresponding Entrez gene IDs using mean values;
Step 3: A rank ordered list of genes was obtained according to the extent of differential expression.

Discovering affected GOMs

Given a query profile, a hyper geometric test is performed for enrichment analysis of every GOM (in default, 2-fold change is used as a threshold to find differentially expressed genes). *P*-values are calculated to indicate if differentially expressed genes are enriched in certain GOMs. GOMs with *P*-value <0.01 are selected. Three basic GO (15) categories [BP (Biological Process), CC (Cellular Component) and MF (Molecular Function)] are provided for comparing GOMs with respect to different biological meanings.

Expression pattern comparison with reference instances

The expression pattern similarities of query and reference profiles in every GOM are calculated to generate a reduced similarity matrix with each column representing an expression profile corresponding to a chemical intervention in reference library and each row representing a GOM enriched in the query profile. The value in each grid of the matrix represents the similarity score between the query and a reference chemical in certain GOM. It is derived based on Kolmogorov–Smirnov statistics and was called connectivity score in Lamb *et al*.'s work (9), here we called it *S* score. After the calculation of *S* score, the *P*-value is calculated to indicate significance of the comparison. Instances with *P*-value <0.05 (in default) are regarded as having significantly similar (with *S* score >0) or reverse (with *S* score <0) pattern of expression with the query in this GOM. Finally, reference instances are ranked decreasingly according to the

number of matched or reverse-matched (*P*-value <0.05) GOMs.

S score calculation. The similarity score is calculated by summarizing Kolmogorov–Smirnov (KS) scores for both over-expressed gene set and under-expressed gene set. We improved the method used in Connectivity Map to make it fit into expression pattern comparison for every single GOM. For each GOM, differentially expressed genes in this GOM are partitioned according to whether they are up- or down-regulated into two groups. KS scores for both up (KS_{up}) and down (KS_{down}) regulated gene groups are calculated, respectively, using a nonparametric rank-based strategy based on Kolmogorov–Smirnov statistics, the procedure is as follows: let t be the number of genes in either the up- or down-regulated gene group and j denote the jth gene according to the rank of differential expression, assuming there are a total of N genes in array, and the position of the jth gene in the rank ordered whole gene list (also ranked according to the extent of differential expression) is $V(j)$, then $KS_{up/down}$ is calculated as follows:

$$a = \max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{N}\right]$$

$$b = \max_{j=1}^{t}\left[\frac{V(j)}{N} - \frac{(j-1)}{t}\right]$$

$$KS_{up/down} = \begin{cases} a, (a > b) \\ -b, (b > a) \end{cases}$$

The KS score calculated using Kolmogorov–Smirnov statistics indicates the extent of similarity of the data distribution of two samples, when applied here it shows whether two profiles have the same pattern of expression. For each GOM, KS_{up} and KS_{down} show, respectively, whether up- and down-regulated genes have the same or reverse pattern of expression between two chemicals. The similarity score (S) for each GOM is finally calculated by integrating KS_{up} and KS_{down} that set S equaling 0 when KS_{up} and KS_{down} have the same algebraic sign and equaling $KS_{up} - KS_{down}$ other wise. Array pair with positive similarity score in a certain GOM means they have similar pattern of expression in this GOM, and vice versa.

P-value calculation. In every run of expression pattern comparison random permutations of genome-wide (GO term based) gene rank is implemented in default 1000 times to calculate 1000 fake S scores and the percent of times that the absolute value of the fake S is larger than the absolute value of real S is the P-value for the real S score.

RESULTS

Gene expression modules-based similarity search

The idea of gene expression modules-based similarity search (GEMS2) is partitioning genes into functionally meaningful categories, forming a module-based Coarse-Graining expression pattern and then performing expression pattern comparison according to the differences within each category (Figure 1). It is composed of the following steps:

Step 1: Establish a library of gene expression profiles corresponding to different chemical interventions. Data from Connectivity Map (build 01) (http://www.broad.mit.edu/cmap_build01/) was used to evaluate our method.

Step 2: Given a query profile, discover significantly affected GOMs using hyper geometric test.

Step 3: Within each GOM, search against the library for profiles having analogous or reverse patterns of expression. A similarity matrix is constructed taking each GOM as a unit, and is summarized to measure the biological distances between query and profiles in library (for a detailed description, see the Materials and Methods section).

Web interface

Based on the GEMS2 method and algorithm, a free web-based service is available to perform online similarity search (http://www.biosino.org/GEMS2/).

Case one: searching for molecules having similar functions

We first demonstrate that our method is efficient in finding chemicals having similar functions. This case comes from a study (21) that investigated the effect of valproic acid (VPA) and all-*trans*-retinoic acid (ATRA) on acute myeloblastic leukemia cells, OCI/AML2. A total of four microarray assays were done in their experiments (data can be downloaded from the gene expression omnibus by ID GDS1215), one array was treated with VPA and another with vehicle. These two were analyzed using our method. After a similarity search, the top 10 chemicals with highest scores were presented (see Table 1). Among them, VPA itself appears three times. For the rest, trichostatin A, vorinostat–HC toxin, though structurally distant are all HDAC inhibitors. Data in the last column (Table 1) shows that these chemicals are almost fully positively correlated with the query, which is consistent with the fact that they perform a similar function. Besides, the cell line used in the query case is myeloblastic leukemia cells which did not exist in our reference library. It indicates that our method is to some extent cell line independent and can provide general functional similarity search among chemical interventions.

Case two: searching for molecules that mimic the cellular response to hypoxia

We then demonstrate our method is also capable of finding chemicals that mimic a certain biological state. This case derives from a work (22) that investigated the effect of hypoxia on 'gene expression' in MCF7 cell line. Six microarray assays in their experiments (three replicates for hypoxia treatment and normoxia treatment, respectively, GDS2758) were analyzed using our method. Search results are presented (Table 2). All top 10 agents show fully positively correlation with the query, and most of them (8 of 10) are reported to have a tight relationship with hypoxia. Among them deferoxamine appears for

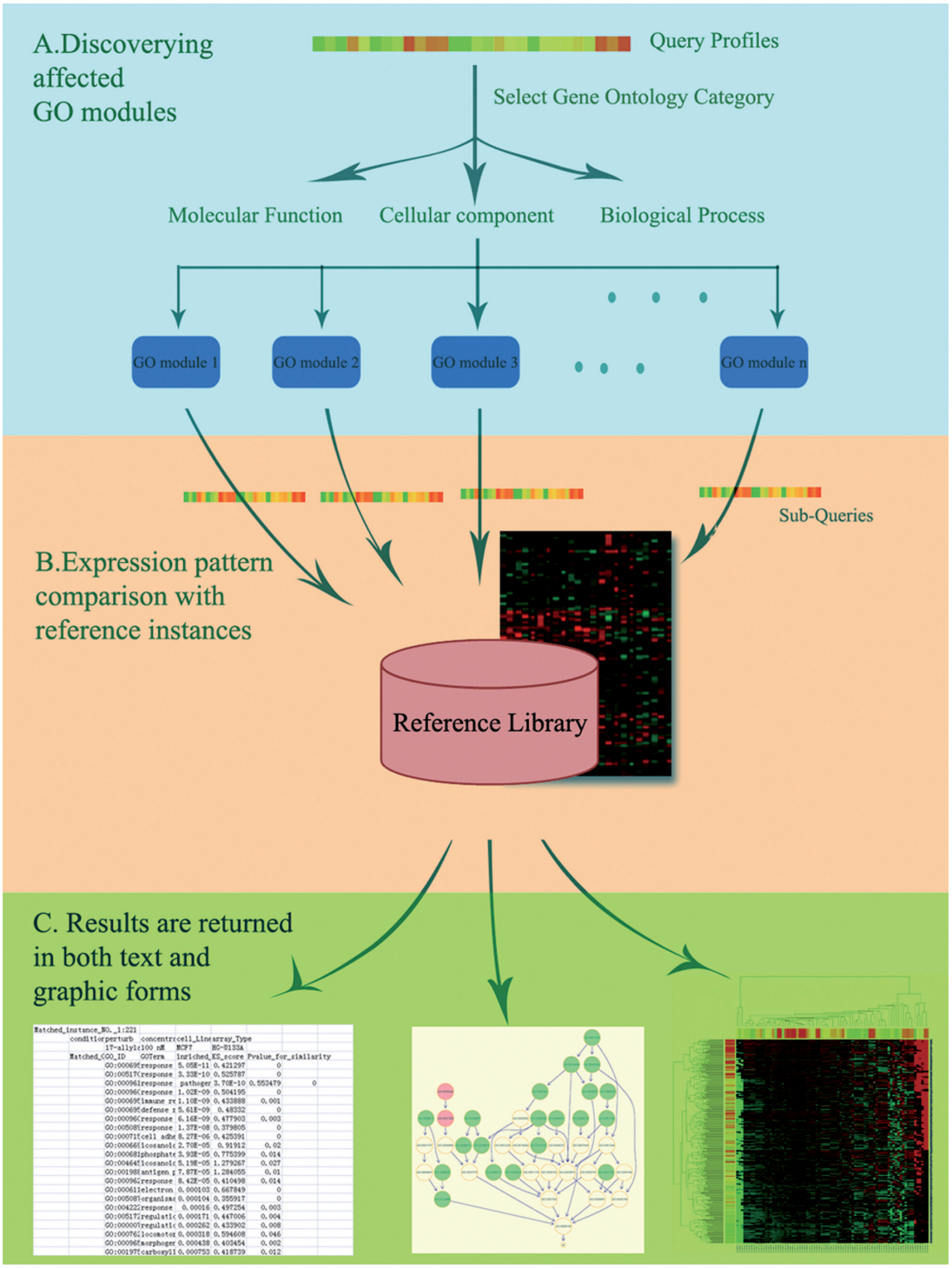

Figure 1. Workflow of gene expression module-based similarity search. Workflow of expression function pattern similarity search: (**A**) A query profile is uploaded and GOMs significantly affected are found using hypergeometric test; (**B**) for each GOM found in the first step, expression pattern comparison is performed between the query and every reference instance to calculate for each GOM a *S* score and a *P*-value. Instances whose *P*-value is above the threshold (default is 0.05) are filtered out. Instances are arranged in descending order according to the number of matched (or reverse-matched) GOMs; (**C**) results are returned in both graphic and textual forms.

three times. Deferoxamine is a chelating agent capable of binding free iron in the bloodstream and removing excess iron from the body. It is usually used as an hypoxia mimic that simulates the hypoxic state by altering the iron status of hydroxylases (23). Dimethyloxalylglycine, a nonspecific inhibitor of 2-OG-dependent dioxygenase, is another hypoxia mimicking agent. Prochlorperazine, though the exact mechanism is unknown, is reported to have the effect of augmenting hypoxic responsiveness in humans (24). Colforsin has the ability of raising levels of cyclic

Table 1. Top 10 reference instances found for profiles of VPA-treated cells

cMap ID	Molecule	Dose	Cell line	GO counts
1072	Trichostatin A	1 μM	MCF7	21(20+, 1−)
410	VPA [INN]	10 mM	HL60	20(20+, 0−)
1000	Vorinostat	10 μM	MCF7	20(20+, 0−)
1050	Trichostatin A	100 nM	MCF7	20(20+, 0−)
909	HC toxin	100 nM	MCF7	19(19+, 0−)
989	VPA [INN]	1 mM	MCF7	19(19+, 0−)
332	Trichostatin A	100 nM	MCF7	18(18+, 0−)
1112	Trichostatin A	100 nM	MCF7	17(17+, 0−)
866	Ikarugamycin	2 μM	MCF7	17(17+, 0−)
409	VPA [INN]	1 mM	HL60	16(16+, 0−)

The top 10 instances sharing the largest number of significantly affected GOMs with VPA-treated cells are listed here. For detailed parameter settings: BP is chosen as GO mode, the permutation time is set to be 1000, and the *P*-value for cutting off insignificantly matched or reverse-matched GO modules is 0.05. 'plus' indicates the number of GO terms positively correlated; 'minus' indicates the number of GO terms negatively correlated.

Table 2. Top 10 reference instances found for profiles of hypoxia treated cells

cMap ID	Molecule	Dose	Cell line	GO counts
573	Deferoxamine [INN]	100 μM	MCF7	57(57+, 0−)
904	5109870	25 μM	MCF7	57(57+, 0−)
584	Dimethyloxalylglycine	1 mM	PC3	52(52+, 0−)
1010	Thioridazine [INN]	10 μM	MCF7	49(49+, 0−)
460	Deferoxamine [INN]	100 μM	PC3	48(48+, 0−)
1053	Prochlorperazine [INN]	10 μM	MCF7	46(46+, 0−)
485	Deferoxamine [INN]	100 μM	MCF7	42(42+, 0−)
977	Wortmannin	1 μM	MCF7	42(42+, 0−)
1001	Sirolimus [INN]	100 nM	MCF7	40(40+, 0−)
913	Colforsin [INN]	50 μM	MCF7	39(39+, 0−)

The top 10 instances sharing the largest number of significantly affected GOMs with hypoxia treated cells are listed here. For detailed parameter settings: GO mode: BP; permutation time: 1000; *P*-value: <0.05.

AMP which leads to the increase of LDH activity. It can also mimic the effects of hypoxia with regard to the hypoxia-induced increase in LDH activity (25). This case demonstrates that our method is quite powerful for finding chemicals that cause or mimic a certain biological state.

But there are two exceptions in these similarity alignments. We know that wortmannin and sirolimus are PI3K and mTOR inhibitors, respectively. Both of them lead to the inhibition of hypoxia-inducible factor's activity (26,27), which supposed to have a reverse effect to hypoxia. But our data (Table 2) shows that cell lines treated by wortmannin and sirolimus have similar enriched GO pattern of expression compared with that of hypoxia treated cell lines. A possible explanation for this is that other mechanisms may exist which result in this kind of positive correlation.

Case three: searching for molecules that reverse the expression pattern of tumorigenic breast cancer cells

In this case we demonstrate that our method can also been used to find novel molecules reversing the effects of disease, which may provide useful information for therapeutics. This case is taken from a work (28) that analyzed expression changes in breast cancer cells having high tumorigenic capacity. Nine microarray assays (three normal and six tumorigenic, GDS2617) were analyzed using our method. Top 10 hits are presented (Table 3). Trichostatin A is histone deacetylase inhibitor, which has long been investigated as a potential antitumor agent against breast cancer (29,30). For the rest, genistein, resveratrol, metformin, novobiocin are also reported to have general antitumor effects (31–36). Data in the last column (Table 3) shows that the effects of all top 10 chemicals are negatively correlated with expression pattern of tumorigenic cells, which is consistent with their antitumor activities. One the other hand, most GOMs found here associated with top 10 chemicals are cell cycle related, which is consistent with the fact that most antitumor chemicals exert their effects directly or indirectly by influencing cell cycle-associated biological processes.

Table 3. Top 10 reference instances found for profiles of tumorigenic breast cancer cells

cMap ID	Molecule	Dose	Cell line	GO counts
448	Trichostatin A	100 nM	PC3	27(6+, 21−)
1015	Genistein	10 μM	MCF7	26(0+, 26−)
841	Resveratrol	10 μM	MCF7	25(0+, 25−)
486	Calmidazolium	5 μM	MCF7	24(0+, 24−)
164	Dexverapamil [INN]	10 μM	MCF7	23(0+, 23−)
2	Metformin [INN]	10 μM	MCF7	23(0+, 23−)
965	Felodipine [INN]	10 μM	MCF7	20(0+, 20−)
435	Novobiocin [INN]	100 μM	PC3	20(0+, 20−)
381	17-allylamino-geldanamycin	1 μM	MCF7	20(19+, 1−)
383	Cobalt chloride	100 μM	MCF7	20(0+, 20−)

The top 10 instances sharing the largest number of significantly affected GOMs with tumorigenic breast cancer cells. For detailed parameter settings: GO mode: BP; permutation time: 1000; *P*-value: <0.05.

Dependency of Connectivity Map on probe number and probe selection

The input of Connectivity Map search system is a small set of rank-ordered up- and down-regulated gene probes. There is no specific restriction for probe number or probe selection. Global similarity search using only a small fraction of genes may cause insufficient information usage, which may lead to instabilities of search result. Here we took the data already used (GDS1215: VPA treatment versus vehicle) to illustrate the problems that may arise when using Connectivity Map improperly. Table 4 shows both chemicals appears in top 10 and their ranks are highly diverse when randomly using top 10, 20 and 30 up- and down-regulated genes as signatures, respectively. This case indicates that in order to get reliable search results by using Connectivity Map, researchers should carefully select gene signatures and may need a step-by-step analysis to find a suitable probe dataset in order to gain a reliable output. On contrary to the Connectivity Map, our methodology groups genes into certain number of GOMs which are dependent only on the structure of GO, and performs a similarity search

Table 4. The result of Connectivity Map is highly dependent on probe number and probe selection

	10 (instance ID/name)	20 (instance ID/name)	30 (instance ID/name)
1	450 (17-Allylamino-geldanamycin)	607 (Butein)	456 (Quinpirole)
2	313 (NU-1025)	456 (Quinpirole)	267 (Genistein)
3	311 (Monastrol)	450 (17-Allylamino-geldanamycin)	410 (Valproic acid)
4	263 (Clofibrate)	410 (VPA)	703 (Genistein)
5	606 (Thalidomide)	317 (*N*-phenylanthranilic acid)	1021 (Estradiol)
6	611 (Geldanamycin)	703 (Genistein)	609 (5666823)
7	868 (5182598)	483 (Imatinib)	332 (Trichostatin A)
8	491 (Dopamine)	413 (Trichostatin A)	508 (Staurosporine)
9	607 (Butein)	1075 (Fluphenazine)	371 (Rofecoxib)
10	1075 (Fluphenazine)	448 (Trichostatin A)	389 (Wortmannin)

Log ratios of VPA-treated versus vehicle-treated gene expression values are calculated. Probes are ranked according to the extent of differential expression. The top 10, 20 and 30 probes up- and down-regulated are picked up. Queries (10 up−10 down, 20 up−20 down, 30 up−30 down) are used to search against connectivity map, respectively. Top 10 instances positively correlated are presented in the table. (probeNum: indicates the number of up/down probes used).

that is just based on the comparison of these GOMs. Therefore, by using our similarity search system it is no longer necessary to consider how to select gene signatures. Furthermore, as shown in case one our strategy is sensitive and can provide reasonable and reliable search results.

DISCUSSION

Algorithms like GSEA (37) and sigPathway (38) have introduced the 'gene set' concept into expression profile analysis and are proved useful in explaining gene expression data. Here we brought this concept into chemical induced expression pattern similarity search, which involves partitioning genes into small biological categories and performing expression pattern comparison within each category.

The main focus of our method consists of two points: first, expression pattern comparison-based chemical function similarity search. Be different from traditional structure comparison that also emphasizes on similarity comparison, expression profile comparison is more straightforward because the rule 'similar structure cause similar function' does not always hold; second, we restricted the similarity comparison into every gene function module, which can not only tell the extent of overall similarity of two chemicals but also can provide information about in which function modules the two chemicals are similar. It can be seen as an improvement of Connectivity Map as it can provide more biological information of the chemicals.

The advantages of this module-based comparison strategy can be summarized in the following three points: first, module-based expression pattern comparison makes it possible to identify in which pathways or functional modules are two profiles similar. This is useful for deducing functions of unknown chemicals more precisely from those of well studied. Second, as shown in our case studies, module-based expression pattern comparison can help us to find chemicals which though structurally distant are functionally alike because they affect similar pathways or biological processes. This advantage will be helpful to detect main or side effects of chemicals or drugs. Third, in our methodology gene expression patterns are reduced into patterns of GOMs, which are depend only on the structure of GO and the similarity search performed is just based on the comparison of these GOMs. There is no longer necessary to consider how to select gene features as done when using Connectivity Map.

Starting from Connectivity Map, some significant improvements were made in our method. First, all of up- and down-regulated genes in the query profile are used to avoid result instability. Concerning the work of Lamb and his colleagues (9), the input of Connectivity Map can be a small set of probes up- and down-regulated. There is no specific rule to restrict probe number, and probe selection is also quite flexible largely depending on the individual researcher's judgment. This has the risk of insufficient information usage and the consequence is that different selection of probe sets may generate diverse and even conflicting outcomes as shown in the results part (Table 4). Because all up- and down-regulated genes (in default using the criteria of 2-fold change) were used to build patterns of GOMs, our method not only avoids insufficient information usage, but also provides a much more stable search result. Second, in Connectivity Map, only the relative similarity score is provided, which can only indicate whether two given chemicals are more similar than two other. There is no way to know to what extent and in which type of functional level two chemicals resemble or whether it is statistically significant. We make up this flaw by using random sampling method (see Materials and Methods section) to calculate a *P*-value for each GOM and give a reasonable evaluation score to chemical similarities.

Finally, although our method can be named as 'similarity searching' approach, it does not focus only on finding the most closely associated chemicals. When search for related profiles for a given chemical, chemicals ranked with higher scores only indicate that they share more GOMs than others ranked with lower scores. It is hard to say that the former is more relevant to the query than the latter, especially when two target chemicals have close ranks. One of the major purposes of our method is providing as much biological information as possible about unknown interventions, which overcomes the limitation that global comparison has.

So far, whole data from Connectivity Map (build 01) was used as a basic library for validating the rationality of our method. Persistent efforts will be made in adding data from other sources, including different disease states, other cell lines and organisms to upgrade and enlarge our data resource, as well as to continue expanding the system's applications. In this article, the GO system is applied as the rule to partition genes. But it does not mean that this rule is superior to others like pathway information, etc. The focus of our article is to propose a new method for expression pattern comparison that combining priori defined gene set information, and the selection of rules on how to partition genes is actually depend on the researchers needs. Till now, our server only provides GO system to define gene set, and we are trying to add more partitioning rules in the server and also considering adding an option for user to upload their user-defined gene sets.

In the case studies our method has shown its power in discovering chemicals sharing similar biological mechanisms and chemicals reversing disease states. Both sides are of great importance in biological and biomedical research, especially for deciphering potential regulatory mechanisms of small molecules on biochemical pathways. Our methodology is also quite useful to help researchers to discover candidate drugs or new usages of old drugs. Furthermore, it may shed light on some applications involved in applied medicine research, such as unknown toxin identification, side effects discovery or prediction and design of disease specific therapeutics.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr Minerva T. Garcia-Barrio for her critical reading of this manuscript.

FUNDING

863 Hi-Tech Program of China (863) (grant 2007AA02Z304, 2007AA02Z332, 2006AA02Z344, 2006AA02Z334, 2006AA020406); the Shanghai Committee of Science and Technology (grant 07dz22004); National Key Basic Research Program (973) (grant 2006CB910700, 2004CB720103, 2004CB518606, 2003CB715901); Research Program of CAS (grant KSCX2-YW-R-112). BMBF CHN07/38 (to H.-J.T.). Research supported by the listed fundings are open for public access.

Conflict of interest statement. None declared.

REFERENCES

1. Wei,Y., Liu,Y., Wang,J., Tao,Y. and Dai,J. (2008) Toxicogenomic analysis of the hepatic effects of perfluorooctanoic acid on rare minnows (Gobiocypris rarus). *Toxicol. Appl. Pharmacol.*, **226**, 285–297.
2. Pogribny,I.P., Bagnyukova,T.V., Tryndyak,V.P., Muskhelishvili,L., Rodriguez-Juarez,R., Kovalchuk,O., Han,T., Fuscoe,J.C., Ross,S.A. and Beland,F.A. (2007) Gene expression profiling reveals underlying molecular mechanisms of the early stages of tamoxifen-induced rat hepatocarcinogenesis. *Toxicol. Appl. Pharmacol.*, **225**, 61–69.
3. Rahman,K.W., Li,Y., Wang,Z., Sarkar,S.H. and Sarkar,F.H. (2006) Gene expression profiling revealed survivin as a target of 3,3′-diindolylmethane-induced cell growth inhibition and apoptosis in breast cancer cells. *Cancer Res.*, **66**, 4952–4960.
4. Lee,S.B., Cha,K.H., Selenge,D., Solongo,A. and Nho,C.W. (2007) The chemopreventive effect of taxifolin is exerted through ARE-dependent gene regulation. *Biol. Pharm. Bull.*, **30**, 1074–1079.
5. Wei,D., Li,M. and Ding,W. (2007) Effect of vanadate on gene expression of the insulin signaling pathway in skeletal muscle of streptozotocin-induced diabetic rats. *J. Biol. Inorg. Chem.*, **12**, 1265–1273.
6. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
7. Fielden,M.R., Brennan,R. and Gollub,J. (2007) A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicol. Sci.*, **99**, 90–100.
8. Nie,A.Y., McMillian,M., Parker,J.B., Leone,A., Bryant,S., Yieh,L., Bittner,A., Nelson,J., Carmen,A., Wan,J. *et al.* (2006) Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Mol. Carcinog.*, **45**, 914–933.
9. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
10. Wei,G., Twomey,D., Lamb,J., Schlis,K., Agarwal,J., Stam,R.W., Opferman,J.T., Sallan,S.E., den Boer,M.L., Pieters,R. *et al.* (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell*, **10**, 331–342.
11. Hieronymus,H., Lamb,J., Ross,K.N., Peng,X.P., Clement,C., Rodina,A., Nieto,M., Du,J., Stegmaier,K., Raj,S.M. *et al.* (2006) Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell*, **10**, 321–330.
12. Segal,E., Friedman,N., Koller,D. and Regev,A. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.
13. Kanehisa,M., Goto,S., Hattori,M., Oki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
14. Krull,M., Pistor,S., Voss,N., Kel,A., Reuter,I., Kronenberg,D., Michael,H., Schwarzer,K., Potapov,A., Choi,C. *et al.* (2006) TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.*, **34**, D546–D551.
15. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
16. Cunliffe,H.E., Ringner,M., Bilke,S., Walker,R.L., Cheung,J.M., Chen,Y. and Meltzer,P.S. (2003) The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. *Cancer Res.*, **63**, 7158–7166.
17. van't Veer,L.J., Dai,H., van,d.V., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der,K.K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
18. Guo,Z., Zhang,T., Li,X., Wang,Q., Xu,J., Yu,H., Zhu,J., Wang,H., Wang,C., Topol,E.J. *et al.* (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinform.*, **6**, 58.
19. Itzkoviz,S., Levitt,R., Kashtan,N., Milo,R., Itzkovitz,M. and Alon,U. (2005) Coarse-graining and self-dissimilarity of complex networks. *Phys. Rev. E.*, **71**, 016127.
20. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**–e15.
21. Trus,M.R., Yang,L., Suarez,S.F., Bordeleau,L., Jurisica,I. and Minden,M.D. (2005) The histone deacetylase inhibitor valproic acid

alters sensitivity towards all trans retinoic acid in acute myeloblastic leukemia cells. *Leukemia*, **19**, 1161–1168.
22. Elvidge,G.P., Glenny,L., Appelhoff,R.J., Ratcliffe,P.J., Ragoussis,J. and Gleadle,J.M. (2006) Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways. *J. Biol. Chem.*, **281**, 15215–15226.
23. Vengellur,A., Phillips,J.M., Hogenesch,J.B. and LaPres,J.J. (2005) Gene expression profiling of hypoxia signaling in human hepatocellular carcinoma cells. *Physiol. Genomics*, **22**, 308–318.
24. Olson,L.G., Hensley,M.J. and Saunders,N.A. (1982) Augmentation of ventilatory response to asphyxia by prochlorperazine in humans. *J. Appl. Physiol.*, **53**, 637–643.
25. Marti,H.H., Jung,H.H., Pfeilschifter,J. and Bauer,C. (1994) Hypoxia and cobalt stimulate lactate dehydrogenase (LDH) activity in vascular smooth muscle cells. *Pflugers Arch.*, **429**, 216–222.
26. Carver,D.J., Gaston,B., Deronde,K. and Palmer,L.A. (2007) Akt-mediated activation of HIF-1 in pulmonary vascular endothelial cells by S-nitrosoglutathione. *Am. J. Respir. Cell Mol. Biol.*, **37**, 255–263.
27. Hudson,C.C., Liu,M., Chiang,G.G., Otterness,D.M., Loomis,D.C., Kaper,F., Giaccia,A.J. and Abraham,R.T. (2002) Regulation of hypoxia-inducible factor 1alpha expression and function by the mammalian target of rapamycin. *Mol. Cell Biol.*, **22**, 7004–7014.
28. Liu,R., Wang,X., Chen,G.Y., Dalerba,P., Gurney,A., Hoey,T., Sherlock,G., Lewicki,J., Shedden,K. and Clarke,M.F. (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N. Engl. J. Med.*, **356**, 217–226.
29. Vigushin,D.M., Ali,S., Pace,P.E., Mirsaidi,N., Ito,K., Adcock,I. and Coombes,R.C. (2001) Trichostatin A is a histone deacetylase inhibitor with potent antitumor activity against breast cancer in vivo. *Clin. Cancer Res.*, **7**, 971–976.
30. Singh,T.R., Shankar,S. and Srivastava,R.K. (2005) HDAC inhibitors enhance the apoptosis-inducing potential of TRAIL in breast carcinoma. *Oncogene*, **24**, 4609–4623.
31. Cappelletti,V., Fioravanti,L., Miodini,P. and Di,F.G. (2000) Genistein blocks breast cancer cells in the G(2)M phase of the cell cycle. *J. Cell Biochem.*, **79**, 594–600.
32. El-Mowafy,A.M. and Alkhalaf,M. (2003) Resveratrol activates adenylyl-cyclase in human breast cancer cells: a novel, estrogen receptor-independent cytostatic mechanism. *Carcinogenesis*, **24**, 869–873.
33. Pozo-Guisado,E., Merino,J.M., Mulero-Navarro,S., Lorenzo-Benayas,M.J., Centeno,F., varez-Barrientos,A. and Fernandez-Salguero,P.M. (2005) Resveratrol-induced apoptosis in MCF-7 human breast cancer cells involves a caspase-independent mechanism with downregulation of Bcl-2 and NF-kappaB. *Int. J. Cancer.*, **115**, 74–84.
34. Dowling,R.J., Zakikhani,M., Fantus,I.G., Pollak,M. and Sonenberg,N. (2007) Metformin inhibits mammalian target of rapamycin-dependent translation initiation in breast cancer cells. *Cancer Res.*, **67**, 10804–10812.
35. Zakikhani,M., Dowling,R., Fantus,I.G., Sonenberg,N. and Pollak,M. (2006) Metformin is an AMP kinase-dependent growth inhibitor for breast cancer cells. *Cancer Res.*, **66**, 10269–10273.
36. Shiozawa,K., Oka,M., Soda,H., Yoshikawa,M., Ikegami,Y., Tsurutani,J., Nakatomi,K., Nakamura,Y., Doi,S., Kitazaki,T. *et al.* (2004) Reversal of breast cancer resistance protein (BCRP/ABCG2)-mediated drug resistance by novobiocin, a coumermycin antibiotic. *Int. J. Cancer*, **108**, 146–151.
37. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
38. Tian,L., Greenberg,S.A., Kong,S.W., Altschuler,J., Kohane,I.S. and Park,P.J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.