Application of connectivity mapping in predictive toxicology based on gene expression similarity

Joshua L. Smalley[1], Timothy W. Gant[1], and Shu-Dong Zhang[1,2]

[1]Medical Research Council Toxicology Unit, University of Leicester, Systems Toxicology Group, Lancaster Road, Leicester, LE1 9HN, UK

[2]Centre for Cancer Research and Cell Biology (CCRCB), Queen's University Belfast, Belfast, UK

Corresponding author: Shu-Dong Zhang (sdz1@le.ac.uk; s.zhang@qub.ac.uk)

Keywords: Connectivity mapping; Query gene signature; Reference gene-expression profile; Predictive toxicology

Abstract

Connectivity mapping is the process of establishing connections between different biological states using gene expression profiles or signatures. There are a number of applications but in toxicology the most pertinent is for understanding mechanisms of toxicity. In its essence the process involves comparing a query gene signature generated as a result of exposure of a biological system to a chemical to those in a database that have been previously derived. In the ideal situation the query gene expression signature is characteristic of the event and will be matched to similar events in the database. Key criteria are therefore the means of choosing the signature to be matched and the means by which the match is made. In this article we explore these concepts with examples applicable to toxicology.

What is connectivity mapping?

The concept of connectivity mapping was first introduced by Lamb et al in 2006 (Lamb *et al*., 2006). It sought to make association between gene expression due to disease state and that due to drug molecules, or similarly between disease state and gene alteration. In making the connections the purpose is to identify molecules which may be used in the treatment of disease, or genes that are the underlying cause. The premise of the methods is that different biological states can be described or characterized adequately using a genomic signature, e.g. the genome-wide mRNA levels as measured by the microarray technologies. The biological connection between two states can then be established by comparing the genomic signatures that represent them based on their signature similarity. In these comparisons similarity is as valuable as diversity. For example if the gene-expression signature of a small-molecule compound is found to be opposite that of a disease state, in other words, the key set of genes are oppositely regulated in the two biological states, then the compound might be used as a drug to treat the disease. Another application of connectivity mapping is to discern properties in a new chemical entity (NCE) by finding positive (or negative for an antagonist) connections with the database of reference compounds. It is this latter application that allows the use of connectivity mapping to make valid predictions on the toxicological properties of a compound if a gene-expression signature for that compound can be obtained. In principal, two compounds may be recognized as having similar toxicological (and/or pharmacological)

properties even if they primarily target different genes of some biological pathway(s) but nevertheless affect a common set of downstream genes similarly.

There are three key components in connectivity mapping: 1. A large collection of pre-built reference gene-expression profiles that serves as a core database, in which each reference profile characterizes a well-defined biological state. 2. A query gene signature that a researcher has compiled as a result of microarray experiments investigating a particular biological condition. 3. A pattern matching algorithm or similarity metric that quantifies the connection between a query gene signature and a reference profile. So a connectivity score is defined as a function between a query gene signature and a reference profile in such a way that it should reflect the underlying biological connections. If the genes in the query signature are similarly regulated in a reference profile, this indicates a strong positive connection between the query gene signature and the reference profile. If, on the other hand, the genes in a signature are oppositely regulated in a reference profile, there is a negative connection between the two. As we already mentioned earlier, a negative connection between a drug-induced gene-expression profile, and a disease gene signature suggest that this drug might be useful to treat the disease.

A connectivity map with practical utility thus must have a core database of reference gene-expression profiles. The first attempt to build such a database was made by researchers in the Broad Institute of MIT and Harvard (http://www.broad.mit.edu/cmap/). In their Connectivity Map Build 01, 164 bioactive small-molecule compounds were applied to 5 selected human cell lines (MCF7, PC3, SKMEL5, HL60, and ssMCF7)

resulting in 564 Affymetrix microarrays, which provided data for 453 individual reference gene-expression profiles. The latest release of Connectivity Map Build 02 of Broad Institute include transcriptomics data of 1309 small-molecule compounds applied to the same 5 selected human cell lines, with 7056 Affymetrix microarrays for 6100 individual reference profiles.

Examples of success

The utility of connectivity mapping was demonstrated previously using some experimentally derived query gene signatures from independent studies, e.g., for HDAC inhibitors (Glaser *et al*., 2003), estrogen (Frasor *et al*., 2004), and immunosuppressive drugs (Horwitz *et al*., 2004). For the HDAC inhibitors query gene signature, vorinostat, trichostatin A, valproic acid, and HC-toxin were found to have significant positive connections to the signature, which were in accordance with the known properties of these compounds as HDAC inhibitors. For the immunosuppressive drugs query gene signature, the identified positive connections include azathioprine (Armstrong *et al*., 2001; Matalon *et al*., 2004), thalidomide (McHugh *et al*., 1995), staurosporine (Ting *et al*., 1995), and trichostatin A (Januchowski *et al*., 2007), the immunosuppressive properties of which were all shown in the corresponding references.

Estrogen gene signature: The Estrogen Receptor (ER) is a group of receptor proteins that are activated by estrogen. The main function of the ER is its gene-regulatory role as a DNA-binding transcription factor. Thus by treating ER positive cells with estrogen, the

ERs are activated to regulate downstream gene expression, and this event results in a gene-expression profile characteristic of the event. We have used an estrogen query gene signature to demonstrate the success of connectivity mapping (Zhang and Gant, 2008). The collection of reference gene-expression profiles were based on the public dataset of the Broad Institute Connectivity Map Build 01 with 453 individual reference gene-expression profiles in total. The query gene signature for estrogen was based on an independent study (Frasor *et al*., 2004) where the MCF7 human breast cancer cell line was treated with the natural estrogen receptor ligand estradiol. The resulting query gene signature included 40 up-regulated and 89 down-regulated genes as selected by the authors of the original study. When the estrogen signature was compared with the database of 453 reference profiles, the following compounds in the database were shown to have significant positive connections with estrogen: Estradiaol, alpha-estradiol, genistein, and NDGA, indicating that these may have similar estrogen receptor binding properties as estrogen. In this example estradiol acted as a test of the method because reference profiles for this compound were known to be present in the database. Therefore the positive connection with estrogen signature was reassuring. The other connections made were also biologically plausible, alpha-estradiol is a stereoisomer of estradiol (Edwards and McGuire, 1980) and genistein is a phyto-estrogen (de Lemos, 2001) from plants. However NDGA was an unexpected finding, at least to us, as we were not aware of NDGA to be an ER related compound. Therefore the question for NDGA was, if it is connected positively with estrogen is there any other support for this in the literature? A text search showed that this was indeed a plausible connection and some references show that NDGA has an estrogenic activity and able to elicit an estrogen-like response

(Fujimoto *et al*., 2004; Sathyamoorthy *et al*., 1994). There were also significant negative connections and these compounds were antagonists of the estrogen receptor. These negatively correlated compounds included the well known estrogen receptor antagonists fulvestrant, tamoxifen, and raloxifene (Buzdar, 2004; Jacobs *et al*., 1988; Fuchs-Young *et al*., 1995).

Pumaprazole gene signature: PredTox is an EU 6th Framework Programme (http://www.innomed-predtox.com) carried out by a consortium of 14 pharmaceutical companies, 3 academic institutions and 2 technology providers. The primary aim of the consortium was to provide the data generated from all the 'omics technologies in addition to results from more conventional toxicology methods, to identify new biomarkers with utility for the early identification of toxicity. In the process it also allowed the development of more effective 'omics related analysis techniques for potential use in drug discovery and development. The samples in the PredTox study were generated from 16 compounds which were tested separately by the partners of the consortium following a common experimental design and treatment schedule. The compounds used included two reference compounds, a hepatotoxin troglitazone and a nephrotoxin, gentamycin. The other 14 compounds were drug molecules that had failed early on in development due to hepatic or nephrotoxic effects. The PredTox database therefore contains toxicological data for 16 compounds tested on rats following a common experimental and treatment design. Transcriptomics profiling data were available for the liver, kidney, and blood samples of each rat.

We first tried to use the liver transcriptomics data for compound FP008AL (pumaprazole) to compile a query gene signature and compared it with the collection of reference profiles based on the Broad Institute Connectivity Map Build 02 dataset. The connectivity mapping exercise using this query gene signature produced significant connections with approximately 100 compounds. The connections made either seemed to be related to disruption of glutathione metabolism or the competition with ATP at ATP binding sites, mainly those of the ATP-binding cassette (ABC) transporters. Lansoprazole was one such compound linked by the latter mechanism.

Lansoprazole is one of the class of drugs known as PPI (proton pump inhibitors), whose main pharmacological action is a pronounced and long-lasting reduction of gastric acid production (Berlin *et al*., 1992). The FP008AL compound in the PredTox study is pumaprazole, a reversible proton pump inhibitor (Martínek *et al*., 1999). The connection between these proton pump inhibitors was a very significant one for verification of the analysis technique, and its significance notable for the following reasons: 1. This was a cross-species comparison, as the query gene signature was based on rat data, while the collection of reference profiles were based on human data. 2. The query gene signature was based on *in-vivo* data, while the reference profiles were *in vitro* data. 3. Unlike the estrogen receptor, the proton pump is not known as a transcription factor. It therefore does not directly regulate transcriptional gene expression as does estrogen receptor. It is therefore all the more significant that a gene expression profile which would have resulted from an indirect effect of the compound on cellular biochemistry has sufficient identity to allow matching with a similar compound across organs, in vivo to in vitro and

species. Therefore despite these 3 apparent barriers to success, the connectivity map was still able to pick up the connection between the pumaprazole query gene signature and the lansoprazole reference profile. This example indicates therefore the power of this analytical method.

Predictive toxicity

Next we tried to use the 16 transcriptomics datasets from the PredTox study to construct reference gene-expression profiles against which to match a query gene signature from another study. If all the data were available there would be 288 individual reference profiles [288 combinations of 16 compounds, 2 doses (low and high apart from vehicle control), 3 time points (day 2, 4, and 15), and 3 tissue types (liver, kidney and blood)]. Some of the data were not available on the PredTox database but with that which were available, we constructed over 250 reference gene-expression profiles each is named according to its derivation, e.g. FP001RO_LowDose_Liver_Day04, where the compound code is first, followed by dose, tissue and time point.

To compare against this collection of PredTox reference profiles, we compiled some query gene signatures from mice treated with the compound griseofulvin. Griseofulvin is an antifungal drug that causes a cholestasis on oral administration due to an irrevsible inhibition of the terminal enzyme of haem biosynthesis in the liver ferrochelatase (Polo *et al*., 1997; Gant *et al*., 2003). This causes an accumulation of protoporphyrin IX in the liver which blocks bile canaliculi leading to cholestasis. In the study (Gant *et al*., 2003) C57BL/6J and BALB/c mice were treated with 1% griseofulvin administered in the diet

over a time course with sampling at days 1, 3, 5, 8, 15 and 21. For each time point there were 4 pairs of mice and microarray analysis was carried out on liver samples from each time point. From the data we compiled 7 query gene signatures using C57BL/6J and BALB/c mice transcription data. We found that on average, each query gene signature has significant connections to 30 reference profiles in the PredTox collection. In this connectivity mapping exercise, the threshold p-value was set such that on average 1 false connection was expected per query signature, so the overall FDR (False Discovery Rate) was estimated as 3.3%. Detailed descriptions of how p-values were calculated and how threshold was set can be found in (Zhang and Gant, 2008). We note here that it is still difficult to effectively estimate the False Negative Rate (FNR), as this will require a power function for the statistical test used in the connectivity mapping exercise, for which statistical advancement is still lagging behind. In this exercise, when queried with the griseofulvin BALB/c day 1 signature, the connectivity map produced statistically significant positive connections with 5 compounds in the PredTox database, and all the reference profiles were liver samples and 3 of which were early time points. Of the 5 compounds found connected to the BALB/c day 1 signature, compounds FP003SE and FP008AL belonged to a subgroup, in which hypertrophy in the liver was the major histopathological finding; and compounds FP005ME and FP014SC belonged to a subgroup, whose classical common findings were bile duct damage, hyperplasia, increased bilirubin and cholestasis, much of the effects listed here were also observed in the griseofulvin-treated BALB/c mice (Gant *et al*., 2003). It is worth noting the following points: 1. An early time point gene signature for griseofulvin is most likely to contain genes associated with its primary biochemical effect because later as pathology

develops much of the differential gene expression is not due to the compound per se but rather the pathophysiological response. 2. The data presented here indicates that, in the case of cholestasis, connectivity analysis could potentially be used to analyse and predict the toxic mechanism of a compound. 3. Significant connections were made despite the query gene signature being mouse derived, and the reference profiles being rat derived.

Software

Implementing the improved methodology of connectivity mapping we introduced in (Zhang and Gant 2008), sscMap is an extensible java application for connecting small-molecule drugs using gene expression signatures (Zhang and Gant 2009). The benefits of the method implemented in this application include a more principled statistical procedure (Tian *et al*., 2005; Efron and Tibshirani 2007; Chen *et al*., 2007), effective safeguards against false connections, and an increased sensitivity. The software is bundled with a default collection of 6100 reference profiles based on the Broad Institute Connectivity Map 02 dataset. It comes with a user-friendly GUI (Graphical User Interface) and detailed tutorial guided instructions for using the program. Users can extend the default collection of reference profiles by adding custom-built reference profiles to sscMap. The software can be freely downloaded from http://purl.oclc.org/NET/sscMap.

Conclusions

Recognition of toxicity at early stage, preferably *in vitro* and if *in vivo* before the development of pathological change is a highly desirable goal which would lead to better

toxicological evaluation at decreased cost. Furthermore the recognition of new drug candidates is also a highly desirable objective. The connectivity map presented here and in previous papers can assist in the achievement of both of these objectives. In drug development it can make connections between molecules according to their pharmacological properties even when there is no direct effect of the compound on gene expression, i.e., the effects on gene expression are secondary to the compound and relate to the altered biological state in the test system. In toxicology the method has applicability for the early recognition of potential toxicity in novel molecules with an indication of mechanism. Therefore applied in the early stages of toxicological evaluation it has the potential not only to identify toxicity in a quantitative manner but also to provide a qualitative assessment to the nature of that toxicity. Furthermore the method is *per se* generic and therefore can be applied to all other data types, in particular proteomics and metabonomics where enough data is generated to produce a signature. The connectivity mapping approach also has a special value in recognizing and predicting similar pharmacology and toxicology in compounds with distant structures. It allows compounds of different chemical structures to be associated if they do have similar pharmacological or toxicological properties. An example was already provided in (Lamb *et al*., 2006) where HC-toxin and valproic acid were identified as HDAC inhibitors despite the fact that they are structurally distant from those HDAC inhibitors used to generate the query signature.

Notwithstanding the success and promise of connectivity mapping, there are a couple of practical issues that need to be addressed before this approach can be widely adopted.

One of them is how we interpret the likely large number of connections picked up by the connectivity mapping exercise, between compounds in a database and a test substance of unknown properties. For example, the PredTox FP008AL compound (pumaprazole) was found to have significant connections to over 100 compounds in the Broad Institute Connectivity Map 02 collections. In that example, lansoprazole was highlighted and discussed as these two compounds were known to have similar pharmacology. The second closely related issue is, with a real unknown compound how do we prioritize the discovered connections and develop new hypotheses, so that we can maximize the efficiencies and success in the following-up efforts? These remain to be open questions, and will be addressed as the development of connectivity mapping approach continues.

Acknowledgements

This work was supported by the Medical Research Council UK. The authors would like to thank all members of the Systems Toxicology Group at the MRC Toxicology Unit for their support, and our collaborators in the FP6 PredTox consortium for access to the PredTox database. SDZ's work at QUB is also supported by the Department for Employment and Learning through its "Strengthening the All-Island Research Base" initiative. We thank the reviewers for their valuable and constructive comments, which helped to improve the manuscript.

Conflict of Interest statement

The authors declare that there are no conflicts of interest.

References

Armstrong, V.W., Oellerich, M., 2001. New developments in the immunosuppressive drug monitoring of cyclosporine, tacrolimus, and azathioprine. Clinical Biochemistry 34, 9-16.

Berlin, I., Molinier, P., Duchier, A., Cournot, A., Durrel, J., Dellatolas, F., Duchier, J., 1992. Dose ranging study of lansoprazole, a new proton pump inhibitor, in patients with high gastric acid secretion. Eur J Clin Pharmacol. 43, 117.

Buzdar, A.U., 2004. Fulvestrant: a new type of estrogen receptor antagonist for the treatment of advanced breast cancer. Drugs Today 40, 751.

Chen, J.J., Lee, T., Delongchamp, R.R., Chen, T., Tsai, C.A., 2007. Significance analysis of groups of genes in expression profiling studies. Bioinformatics 23, 2104-2112.

de Lemos, M.L., 2001. Effects of soy phytoestrogens genistein and daidzein on breast cancer growth. Ann Pharmacother. 35, 1118.

Edwards, D.P., 1980. 17 ALPHA-ESTRADIOL IS A BIOLOGICALLY-ACTIVE ESTROGEN IN HUMAN-BREAST CANCER-CELLS IN TISSUE-CULTURE. ENDOCRINOLOGY 107, 884.

Efron, B., Tibshirani, R., 2007. On testing the significance of sets of genes. Ann Appl Statist 1, 107-129.

Frasor, J., Stossi, F., Danes, J.M., Komm, B., Lyttle, C.R., Katzenellenbogen, B.S., 2004. Selective Estrogen Receptor Modulators: Discrimination of Agonistic versus Antagonistic Activities by Gene Expression Profiling in Breast Cancer Cells. Cancer Res 64, 1522-1533.

Fujimoto, N., Kohta, R., Kitamura, S., Honda, H., 2004. Estrogenic activity of an antioxidant, nordihydroguaiaretic acid (NDGA). Life Sciences 74, 1417-1425.

Fuchs-Young, R., Glasebrook, A.L., Short, L.L., Draper, M.W., Rippy, M.K., Cole, H.W., Magee, D.E., Termine, J.D., Bryant, H.U., 1995. Raloxifene is a tissue-selective agonist/antagonist that functions through the estrogen receptor. Ann N Y Acad Sci. 761, 355.

Gant, T.W., Baus, P.R., Clothier, B., Riley, J., Davies, R., Judah, D.J., Edwards, R.E., George, E., Greaves, P., Smith, A.G., 2003. Gene expression profiles associated with inflammation, fibrosis, and cholestasis in mouse liver after griseofulvin. EHP Toxicogenomics 111(1T), 37.

Glaser, K.B., Staver, M.J., Waring, J.F., Stender, J., Ulrich, R.G., Davidsen, S.K., 2003. Gene Expression Profiling of Multiple Histone Deacetylase (HDAC) Inhibitors: Defining a Common Gene Set Produced by HDAC Inhibition in T24 and MDA Carcinoma Cell Lines. Mol Cancer Ther. 2, 151-163.

Horwitz, P.A,, Tsai, E.J., Putt, M.E., Gilmore, J.M,, Lepore, J.J., Parmacek, M.S., Kao, A.C., Desai, S.S., Goldberg, L.R., Brozena, S.C., Jessup, M.L., Epstein, J.A., Cappola, T.P., 2004. Detection of Cardiac Allograft Rejection and Response to Immunosuppressive Therapy With Peripheral Blood Gene Expression. Circulation 110, 3815-3821.

Jacobs, A.L., Edgerton, L.A., Silvia, W.J., Schillo, K.K., 1988. Effect of an estrogen antagonist (tamoxifen) on cloprostenol-induced luteolysis in heifers. J Anim Sci. 66, 735.

Januchowski, R., Jagodzinski, P.P., 2007. Trichostatin A down-regulates ZAP-70, LAT and SLP-76 content in Jurkat T cells. International Immunopharmacology 7, 198-204.

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., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R., 2006. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science 313, 1929-1935.

Martínek, J., Blum, A.L., Stolte, M., Hartmann, M., Verdú, E.F., Lühmann, R., Dorta, G., Wiesel, P., 1999. Effects of pumaprazole (BY841), a novel reversible proton pump antagonist, and of omeprazole, on intragastric acidity before and after cure of Helicobacter pylori infection. Aliment Pharmacol Ther. 13, 27.

Matalon, S.T., Ornoy, A., Lishner, M., 2004. Review of the potential effects of three commonly used antineoplastic and immunosuppressive drugs (cyclophosphamide, azathioprine, doxorubicin on the embryo and placenta). Reproductive Toxicology 18, 219-230.

McHugh, S., Rifkin, I., Deighton, J., Wilson, A., Lachmann, P., Lockwood, C., Ewan, P., 1995. The immunosuppressive drug thalidomide induces T helper cell type 2 (Th2) and concomitantly inhibits Th1 cytokine production in mitogen- and antigen-stimulated human peripheral blood mononuclear cell cultures. Clin. Exp. Immunol. 99, 160-167.

Polo, C.F., Buzaleh, A.M., Vazquez, E.S., Afonso, S.G., Navone, N.M., Batlle, A.M., 1997. Griseofulvin-induced hepatopathy due to abnormalities in heme pathway. Gen Pharmacol. 29, 207.

Sathyamoorthy, N., Wang, T., Phang, J., 1994. Stimulation of pS2 expression by diet-derived compounds. Cancer Res 54, 957-961.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., Park, P.J., 2005. Discovering statistically significant pathways in expression profiling studies. PNAS 102, 13544-13549.

Ting, C.C., Wang, J., Hargrove, M.E., 1995. Reversal of multiple-site tumor cell-induced immunosuppression by specific cytokines and pharmacological agents. Immunopharmacology 30, 119-130.

Zhang, S.D., Gant, T.W., 2008. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. BMC Bioinformatics 9:258.

Zhang, S.D., Gant, T.W., 2009. sscMap: An extensible Java application for connecting small-molecule drugs using gene-expression signatures. BMC Bioinformatics 10:236.