Using GO ontology to improve Connectivity Map connections

Max Libbrecht*, Joel Dudley†

April 4, 2010

Abstract

Drug discovery is an extremely expensive process due to the cost of running tests on animal models and the large number of such tests required. Recently, researchers have had success using computational models to target drug testing, drastically decreasing the number of trials necessary for each successful treatment found. In particular, the Connectivity Map has had success with an approach based on comparing gene expression signatures of drugs and diseases. Their goal, and that of this work, is to predict likely candidate drugs given a set of genes involved in a particular disease. We sought to improve upon the Connectivity Map by associating genes with GO terms, and thereby widening the scope of the gene set from a few proteins to a subset of the biology influenced by the disease. We tested our method on Type 2 Diabetes, with accuracy similar to that of the Connectivity Map. We also attempted to use the same method to predict drug mechanism, but did not find significant results.

1 Introduction

Finding drugs to treat disease is a very expensive process. Traditionally, drugs have been tested with very little direction and therefore requires very many iterations before an effective treatment is found. Furthermore, drug testing is generally performed on animal models close to humans (mice in particular), making the process very expensive.

Computational methods show great promise towards reducing this expense. Already, researchers are using computational methods to direct the testing of new drugs and finding side effects. In the future, the cost associated with finding drug treatments may be significantly reduced, as drugs are increasingly tested effectively *in silico* rather than *in vivo*.

2 Previous work

Due to the economic need for a effective means of finding drug treatments, a great deal of work has been done on this problem. Many approaches have been used, including those based on drug-ligand binding affinity [Grosdidier et al., 2007], and gene expression signature. In the latter category, the Connectivity Map used gene expression signatures of cell lines treated with various drugs to predict candidate drugs given a set of genes influenced by a disease.

2.1 Connectivity Map

Since our work is an extension of the Connectivity Map and we use the Connectivity Map data, we will go into some detail on the project. The Connectivity Map produced a set of gene expression profiles by treating cell lines each with one of 6100 drugs respectively. They took the gene expression profile from each of these experiments and ranked the genes from most over-expressed to most under-expressed. They published this in an online tool

*maxl@cs.stanford.edu
†jdudley@stanford.edu

Figure 1: A pictorial demonstration of the computation of the KS score. Let the grey bar represent a ranking of all the genes, from most over-expressed to most under-expressed in the presence of a particular drug. Given a set of genes, such as those associated with a GO term, we can highlight those genes to produce a "barcode" like this. Intuitively, if the highlighted genes cluster towards the top as they do here, it's likely that there exists a connection between the drug and the GO term.

that suggests the similarity between a set of genes and a drug using the Kolmogorov-Smirnov statistic. They also published the data, which we use as the basis for our method. [Lamb et al., 2006]

3 Methods

While these methods have been largely successful, all methods based directly on a small set of target genes have the disadvantage that they miss a great deal of the metabolic picture. Even though diseases are usually characterized by a small set of genes, they generally influence large subsets of biology of the cells they inflict. Therefore a method that takes into account the larger scale influence of the disease may be able to improve on the other methods. Gene relationship systems, such as the GO gene ontology or the KEGG pathway database could be used to translate from a small set of genes to the larger metabolic picture. We chose to use GO term enrichment as a model for the influence of these genes. (In particular, we used the Biological Process subset of GO terms, which we felt were most biologically relevant and which we found to produced the best results.)

3.1 Drug-GO similarity

To relate drugs and GO terms, there must be a metric of similarity between the two. GO terms are characterized by an associated set of genes, so to compare drugs to GO terms, an association between drugs and genes is needed. The Connectivity Map data provides such an association. As described above, for each drug, the Connectivity Map produced a ranking of genes, from most up-regulated by the drug to most down-regulated. To compare a set of genes with a ranking, the Kolmogorov-Smirnov (KS) statistic is generally used, in particular by GSEA [Subramanian et al., 2007] and the Connectivity Map. We adopted this approach as well.

To compute the KS score between a drug and a GO term, we list all of the indices of the genes associated with the GO term in the gene ranking associated with the drug. Intuitively, if most of the genes fall near the top of the ranking, it's likely that the drug and the GO term have a positive association, and inversely so if most of the genes fall near the bottom of the ranking. See Figure 1 for a pictorial demonstration of this. This intuition is formalized as the KS score, which is close to 1 for positive relationships and close to -1 for negative relationships. By computing the KS score between every drug-GO term pair, we construct a matrix of the scores for every pair. While the KS score can be positive or negative, it's unclear which is more significant, so we took the absolute value of all the KS scores.

3.2 Ranking drug candiates

After computing the matrix of drug-GO term similarity, we can use it to find drug candidates. Given a set of GO terms associated with a disease, we can select the rows in the matrix associated with the GO terms and average

Our method														
metformin	17	171	350	459	617	689	1444	1784	4440	4549				
rosiglitazone	801	1009	1090	1180	1256	3324	3398	3569	3636	4239	4316	4471	5177	5342
Connectivity Map														
metformin	104	202	398	526	734	1253	1953	3074	3396	4023				
rosiglitazone	507	660	1944	2795	3342	3371	3659	4154	4547	4796	4869	4902	5115	5588

Table 1: Results of running our algorithm on gold-standard diabetes genes, in comparison to Connectivity Map. The rows are lists of indices into the entire ranking of drugs. The Connectivity Map data includes multiple replicates for each drug. Each index represents the position where one replicate fell in the drug ranking. There are 6100 drugs total in the Connectivity Map data. Our method turned up metformin at a very high rank. Had metformin been unknown as a diabetes drug, testing the top 20 (say) drugs returned by our method would have uncovered it.

them in some way to produce a score for ever drug. The top scoring drugs are likely to be good candidates for treatments to the query disease.

There are several methods that could be use to combine the scores in the rows of the matrix corresponding to each GO term. We tried using both the mean and the max (corresponding to the L_1 and L_∞ norms), and found that performance was best using the mean.

4 Results

Testing our method is best done on a disease that has established drug treatments and a set of well-supported genes associated with it. We chose type 2 Diabetes, which has both of those qualities. We plugged a set of gold-standard type 2 Diabetes genes [Kodama, 2010] into our algorithm, hoping that existing drug treatments would be highly ranked in the output. In addition to the gold-standard genes, we tested our method on results directly from a gene expression microarray to see how resistant our method is to noise.

The Connectivity Map has data on two diabetes treatments: metformin and rosiglitazone. For both drugs, the data contains a number of replicates of the same experiment. Instead of trying to merge these data, we used the separate measurements as independent tests of our method's performance.

We will display our results as a list of indices into the entire ranking of drugs. For example, suppose metformin replicates appeared at positions 20, 50 and 100 respectively. We would display this "metformin: 20, 50, 100." There are 6100 drugs total (including replicates) in the Connectivity Map data.

4.1 Gold-standard genes

We obtained a set of 20 gold-standard type 2 Diabetes genes and performed GO term enrichment on this set. We plugged the top 10 most enriched GO terms into our algorithm. For comparison, we plugged the genes themselves into the Connectivity Map. The results are shown in Table 1. Notably, metformin appeared at position 17, well above its first occurrence in the Connectivity Map results. The metformin sequence has p-value $p = .01167$, indicating that the ranking is not a product of random chance. On the other hand, the rankings of the other metformin replicates and the performances of the microarray-derived test indicate that degree of the improvement over the Connectivity Map results are likely largely random.

We also tested our method on a set of genes derived from a gene expression microarray experiment. We obtained the results from a microarray from GEO [Edgar et al., 2002], accession number GSE15653 [Pihlajamaki et al., 2009]. This experiment took data from 5 obese diabetics and 5 lean controls. We used SAM [Tusher et al., 2008] to test each gene for significant over-expression. It turned out that the lowest FDR for the genes in the experiment was 13%, well above the standard cutoff of 5%. This does not come as a large surprise, since the experiment was done in human patients, which is a noisy environment. Still, we picked the 60 most over-expressed

Our method														
metformin	96	739	1496	1974	2101	2132	2557	2616	2710	3491				
rosiglitazone	1075	1183	1722	1849	1967	1976	2271	2428	3444	3716	4218	4735	4948	5333
Connectivity Map														
metformin	96	344	564	1243	1691	1854	2604	3924	4478	5408				
rosiglitazone	351	1041	1092	1228	1234	1288	1342	2240	4545	4644	4804	4833	5550	5916

Table 2: Results of running our method on expression array derived genes, in comparison to Connectivity Map. See Table 1 for information on how to read the table. Our method performed slightly worse than Connectivity Map on these genes.

genes to use for the next step.

We then performed GO enrichment on the 60 genes selected from the microarray experiment. Somewhat surprisingly given the noise in the process that generated it, many GO terms were significantly enriched in this set, even after Bonferroni correction. As above, we chose the top 10 enriched GO terms and plugged them into our algorithm, as well as plugging the genes into Connectivity Map for comparison. The results can be seen in Table 2.

5 Determining drug mechanism

We can also use our method to answer another, different question: Given a drug, what GO terms are most strongly associated with it? This may turn up drug mechanisms or side effects. This is an easy question to ask, because methodologically it is the exact opposite question of the previous one. However, we did not find significant results here, so we will not spend very long discussing it.

We can perform this query by selecting the row in the drug-GO term similarity matrix that corresponds to the drug in question. This produces a scoring of GO terms' similarity to the given drug. The top scoring GO terms could be expected to indicate the mechanism of the drug in question.

Our results on diabetes drugs can be seen in Table 3. While this is an interesting question, our method turns out not to be useful in answering it. It shouldn't come as a great surprise that this method performs poorly on this problem, since determining drug mechanism has traditionally been very difficult, and our method is a rather naive approach.

6 Future work

While the results of method are comparable to Connectivity Map, it fails to provide a significant improvement. However, this is just one example of mapping genes to another space where diseases and drugs might have a more natural interface. If such a space exists, this mapping might provide a vastly better drug screening method than one built on genes alone. We plan to perform a similar analysis on KEGG pathways to see if this produces better results than the one built on GO ontology. In the future, another system of gene similarity may provide a more accurate predictor yet.

References

[Edgar et al., 2002] Edgar, R., Domrachev, M., and Lash, A. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207.

GO term	description
metformin 1	
GO:0045103	intermediate filament-based process
GO:0034063	stress granule assembly
GO:0033962	cytoplasmic mRNA processing body assembly
GO:0031109	microtubule polymerization or depolymerization
GO:0070286	axonemal dynein complex assembly
metformin 2	
GO:0051154	negative regulation of striated muscle cell differentiation
GO:0035104	positive regulation of transcription via sterol regulatory element binding
GO:0006760	folic acid and derivative metabolic process
GO:0008292	acetylcholine biosynthetic process
GO:0043923	positive regulation by host of viral transcription
rosiglitazone	
GO:0060307	regulation of ventricular cardiomyocyte membrane repolarization
GO:0060299	negative regulation of sarcomere organization
GO:0033292	T-tubule organization
GO:0014819	regulation of skeletal muscle contraction
GO:0045355	negative regulation of interferon-alpha biosynthetic process

Table 3: Results from using our method to attempt to determine drug mechanism. The headings metformin 1 and 2 refer to two replicates of metformin data. By our analysis, the GO terms returned don't bear any resemblance to the true mechanism of either drug. In addition, the drastically different results from the two replicates of metformin show that this method has high variance as well as being inaccurate. There is one tantalizing association: rosiglitazone is associated with many GO terms related to the heart muscle, and rosiglitazone is known to cause heart disease in some cases as a side effect. However, heart disease is usually caused by blood clots, not by failure of the heart muscle, so it is unlikely that the results returned by our algorithm represent a real association.

[Grosdidier et al., 2007] Grosdidier, A., Zoete, V., and Michielin, O. (2007). EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins: Structure, Function, and Bioinformatics*, 67(4).

[Kodama, 2010] Kodama, K. (2010). Publication pending.

[Lamb et al., 2006] Lamb, J., Crawford, E., Peck, D., Modell, J., Blat, I., Wrobel, M., Lerner, J., Brunet, J., Subramanian, A., Ross, K., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929.

[Pihlajamaki et al., 2009] Pihlajamaki, J., Boes, T., Kim, E., Dearie, F., Kim, B., Schroeder, J., Mun, E., Nasser, I., Park, P., Bianco, A., et al. (2009). Thyroid Hormone-Related Regulation of Gene Expression in Human Fatty Liver. *Journal of Clinical Endocrinology & Metabolism*, 94(9):3521.

[Subramanian et al., 2007] Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, 23(23):3251.

[Tusher et al., 2008] Tusher, V., Tibshirani, R., and Chu, G. (2008). Significance analysis of microarrays. US Patent 7,363,165.