CLUSTERING GENE EXPRESSION PROFILES VIA PERMUTATION VECTORS

STEPHEN C. BILLUPS, LANCE LANA* AND PAULINE GEE†

Abstract. A promising technology for drug discovery is the use of gene expression profiles for lead compound optimization. The central idea of this technology is to use gene expression profiles as a means of screening new drug leads for further development. The key to this screening is to cluster the gene expression profiles of known drugs, so that compounds with 'similar' profiles can be identified. In this paper, we propose a new approach for representing gene expression profiles using permutation vectors.

Key words. gene expression profiles, clustering

1. Introduction. Gene expression profiling is based on measuring gene expression levels in response to various doses of a pharmaceutical compound. Each gene in a cell produces a product, usually protein, but always through an RNA intermediate, that alters the activity of the cell. Often the rate at which this product is created varies according to the environment. We call the amount of product at a given time the *expression level* of the gene. A *gene expression profile* of a drug consists of the expression levels of various genes at a variety of different concentrations of the drug. Such a profile is illustrated in Figure 1.1. The height of each bar in the figure represents the expression level for a particular gene at a particular dose of the drug.

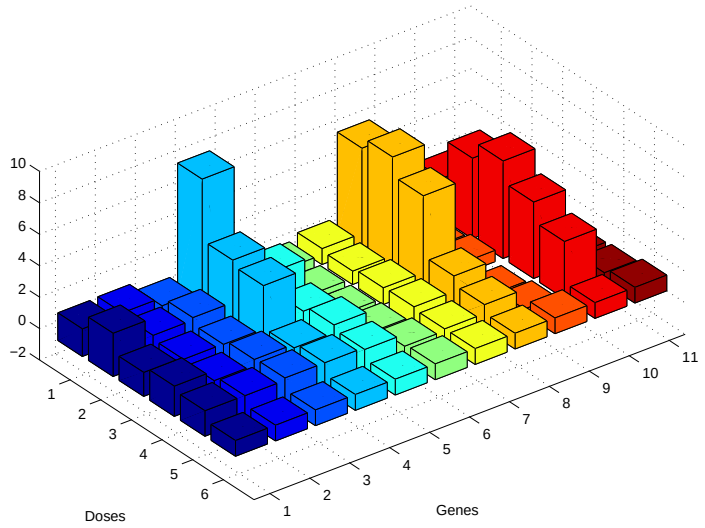

FIG. 1.1. *Sample gene expression profile*

Gene expression profiling represents a promising tool for drug discovery. By analyzing the gene expression profiles of known drugs, criteria can be developed by

*Department of Mathematics, University of Colorado, Denver, Colorado 80217-3364.
†Xenometrix, Inc., a wholely-owned subsidiary of Discovery Partners International, Inc., Boulder, Colorado 80301-5700

which candidate drugs (called leads) can be ranked, so that only the most promising ones are considered for further development and testing.

One approach to analyzing gene expression profiles is to apply clustering methods to group the gene expression profiles in a database of known drugs. Given these groupings, a new lead compound could be evaluated by measuring how close its gene expression profile is to any of the identified clusters. In theory, if a new lead is close to an existing cluster, it is likely to be similar in biological effect. In contrast, leads that are unlike any of the clusters are unlikely to produce similar biological effects.

For this approach to be effective, it is essential to find a representation of the gene expression profiles so that the distance between profiles correlates well with biological effects. One difficulty in establishing such a representation is that gene expression databases often include heterogeneous data. In particular, the concentrations used for one drug may be vastly different than the concentrations used for another drug. Another difficulty, is that, even if identical drug concentrations are used, different drugs may be effective at different concentrations. The representation of the gene expression data, therefore, needs to be insensitive to scaling of the concentrations.

One data representation that has previously been considered is the use of induction concentration (IC) levels [1]. The idea is to determine a dose for each drug that corresponds to the same level of cell mortality. For example, the dose might be chosen so that the population of cells in the sample has decreased by 15% compared to a sample with no drug present. This dose will be different for each drug, but the idea is that there is a common biological meaning across drugs for the doses chosen. Once this dose is determined, gene activities are interpolated for that dose. These interpolated activities for each of the genes are then used to represent the gene expression profile for the drug. If a more detailed representation of the data is needed, then the gene activities at several different IC levels can be used.

This approach has several weaknesses. First, the connection between cell mortality and drug activity is not well understood, and is tenuous at best. Second, some drugs of interest result in an increase in cell population rather than a decrease. These drugs would therefore be excluded from the analysis. Third, it can sometimes be difficult to collect data for drugs at sufficiently high concentrations to reach the desired cell mortality.

In this paper, we propose a new approach for representing gene expression profiles. The approach is based on determining the order (relative to drug concentrations) that the genes are activated. Each drug is then represented by a vector of integers, called a *permutation vector*, which represents the order of the gene activation. Associated with this representation is a distance measure that tells us how similar two permutation vectors are. The details of this approach are discussed in Section 2, and conclusions and future work are discussed in Section 3.

2. Permutation Approach. When the expression level of a gene is affected significantly by the presence of a drug, we say that the gene has been "turned on" by the drug. The permutation approach works on the assumption that drugs that turn on the same genes are similar to one another. Further, drugs that turn on the same genes in the same order (as drug concentrations are increased) are even more similar. This assumption can be rationalized by the following simplistic argument. We can

think of the genes that are affected by a drug as being of two types: beneficial genes, which yield some positive effect; and toxic genes, which indicate toxic levels of the drug. If the toxic genes are turned on at a lower concentration than the beneficial genes, then the drug is not usable because the beneficial effects are not realizable at safe doses. In contrast, if the beneficial genes are turned on at lower concentrations, then there is a potential therapeutic range–that is a safe concentration that yields a beneficial effect.

The implementation of this approach involves four steps. First we define what it means for a gene to be turned on. Then, we define a permutation vector for each drug that gives the order that the drug turned on each of the genes. Once we know the order in which each drug activates genes, we can compare drugs and establish a measure of similarity between them. Using the drug similarities, we are then able to cluster the data using any of a wide variety of clustering algorithms.

Gene Activation - When is a gene "turned on"?. A threshold value must be defined to determine when a gene is first turned on. How best to define this threshold value is a topic of further research. For this paper, we define a gene to be *turned on* when its expression level is at least one standard deviation above the mean gene response, measured over the entire database of known drugs.

Permutation Vectors - Ordering gene response:. Each of the genes is assigned a number 1 to n, where n is the number of genes represented in the database. For each drug, a vector is created recording the order in which the genes are turned on by the drug as drug concentrations increase. If a gene expression never reaches its threshold value at any measured concentration of the drug, it does not appear in the permutation vector for the drug. The permutation vector is padded with zeros at the end to ensure that all vectors have the same length.

As an example, consider the following two permutation vectors for 5 genes:

$$A = \begin{pmatrix} 3 & 4 & 5 & 1 & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 3 & 4 & 2 & 5 & 0 \end{pmatrix}$$

The permutation vector for drug A indicates that drug A turns on genes 3, 4, 5, and 1, in that order. Gene 2 is not turned on by drug A. A zero is used as a placeholder for genes that are not turned on. Similarly, drug B turns on genes 3, 4, 2, and 5 in that order.

Distance measure:. We define a distance measure that attempts to account for differences in the relative orders of genes. This is accomplished by creating a precedence matrix for each permutation vector. The (i, j)th entry of this matrix is set to 1 if gene j appears no later than gene i in the permutation vector. In the example above, the precedence matrices for drugs A and B are given below.

Precedence Matrix - Drug A:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Precedence Matrix - Drug B:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Notice that if gene i appears in the permutation vector, then the ith diagonal entry of the matrix is 1. This ensures that the distance between different permutation vectors is always positive.

Once the precedence matrices have been calculated, a distance measure could be defined simply by adding up the number of entries that are different between precedence matrices. For example, the two matrices above differ in 8 locations, so the distance between them would be 8. Such an approach, however, suffers from a significant bias: drugs that turn on many genes would tend to be further apart than drugs that involve only a few. To illustrate, consider the following two pairs of permutation vectors:

$$A = [1\ 5\ 0\ 0\ 0\ ] \qquad A' = [1\ 2\ 3\ 4\ 5\ ]$$
$$B = [5\ 1\ 0\ 0\ 0\ ] \qquad B' = [5\ 2\ 3\ 4\ 1\ ]$$

Using the above distance measure, the distance between A and B would be 2, whereas the distance between A' and B' would be 14. This large disparity between distances is inconsistent with the fact that in both pairs, the only difference between the vectors is that genes 1 and 5 have been reversed.

To correct this bias, we calculate the distances according the the formula

$$\text{distance} = \frac{N}{M},$$

where N is the number of entries in the precedence matrices that differ, and M is the number of entries where the matrices could have differed in that a one appeared in that entry in at least one of the matrices. For the example above, the distance between A and B is $2/4 = 1/2$, and the distance between A' and B' is $14/22 = 7/11$.

Notice that the distance between two drugs will vary between zero and one. If two drugs turn on the exact same genes in the exact same order, they will have a distance of zero. If two drugs turn on entirely different genes, then their distance will be one. As the distances between different drugs are calculated, they are stored in a distance matrix. With this distance matrix, it is then possible to use various cluster analysis routines to cluster the data.

Clustering Methods:. Having defined a distance measure between permutation vectors, we can perform cluster analysis using any clustering method that relies solely upon distances. For this report, we use hierarchical clustering [2]. Hierarchical clustering works by first assigning every data element to its own cluster, and then iteratively merging clusters together according to some measure of the distance between the clusters, with the clusters that are closest together being merged first. The output of hierarchical clustering is a dendrogram, which is a graphical representation of how the data are merged together. A sample dendrogram is shown in Figure 2.1. Each horizontal bar in the dendrogram represents the merger of two clusters, which are represented by the two vertical bars that are connected by the horizontal bar. The height of the horizontal bar represents the distance between the two clusters being merged.

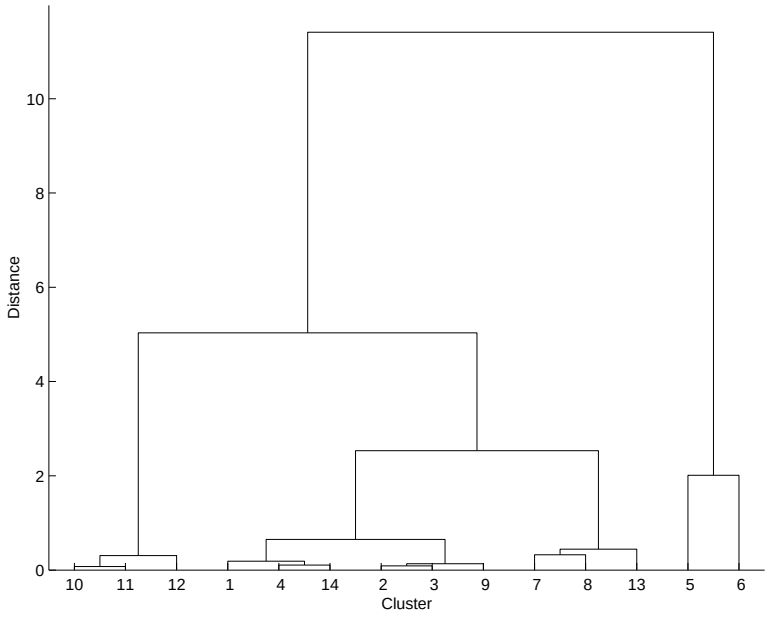

FIG. 2.1. *Sample dendrogram*

The distance between clusters is defined by two components: the distances between individual data elements, and a *linkage method.* The linkage method determines how distances between individual pairs of data are combined to define a distance between clusters. Some common linkage methods are as follows:

- Single linkage: the distance between two clusters is defined to be the shortest distance between an element of one cluster and an element of the other cluster.
- Complete linkage: the distance between two clusters is the largest distance between an element of one cluster and an element of the other cluster.
- Average linkage: the distance between two clusters is the average over all possible pairs of elements with one element from each cluster.
- Ward's method: minimizes within-cluster sum of squares.

3. Conclusions and Future Work. The permutation vector approach presented in this paper is an idea in its infancy. At this stage, it is difficult to assess the

quality of the method. The approach has been used to cluster a proprietary database of gene expression profiles from Xenometrix, Inc. The significance of the resulting clusters is still being evaluated and cannot be reported here.

Perhaps the biggest weakness in the approach is its sensitivity to small perturbations in the data. Indeed, if several genes are turned on at nearly the same drug concentration, then a small change in the data could result in very different orderings of these genes in the permutation vectors. This can result in large distances between very similar drugs. One possible approach for lessening this sensitivity would be to represent the activation order using a probabilistic precedence matrix instead of a permutation vector. The (i, j)th entry of the matrix would represent the probability (given the data) that gene i is turned on before gene j.

Another issue that needs to be addressed is how best to define the thresholds for deciding when a gene is "turned on". Our current approach (a one standard deviation change in the expression level) is arbitrary. Something more reasonable might become apparent from a more careful examination of the data.

Finally, the method provides no mechanism for genes turning off at higher doses. Enriching the representation to account for such events remains a topic for future research.

4. Acknowledgements. The ideas in this paper grew out of a Mathematics Clinic conducted at the University of Colorado at Denver and sponsored by Xenometrix, Inc. We a grateful to all of the students who participated in this clinic. They are Raphael Bar-Or, Joseph Burnham, Jason Gannon, Yvonne Garcia, Supreet Kaur, Lance Lana, and Cary Miller. Other results from the clinic are detailed in the clinic report [1].

REFERENCES

[1] S. C. Billups, R. Bar-Or, J. Burnham, J. Gannon, Y. Garcia, S. Kaur, L. Lana, and C. Miller. Final report of the CU-Denver Mathematics Clinic: Lead compound optimization using gene expression profiling. Mathematics Clinic Report MC01S001, University of Colorado at Denver, Department of Mathematics, Spring Semester 2001.

[2] B.S.Everett. *Cluster Analysis.* Edward Arnold, New York, 1993.