

## A Statistical Method for Identifying Modes of Action

In 'An Information-Intensive Approach to the Molecular Pharmacology of Cancer' (Science, 275 pp 343-349), the authors describe a method of analyzing data collected on more than 60,000 compounds tested against 60 human cancer cell lines. In that approach, measures of the possible modes of action are made for each cell and compared with the activity of each compound in order to determine the modes of action of each compound. Part of that method involves a linear ordering of compounds and targets which is achieved through cluster analysis. In this paper, a method is discussed which aims to identify distinct modes of action appearing in the compound activity data, and thereby cluster compounds appearing to have the same mode of action.

Each compound in the data set is tested against a number of cells. For each cell, the concentration of the compound which reduces the natural growth rate by 50% is determined, this is labelled GI50 in the data set and will be referred to simply as the I50 here, its logarithm will be denoted by L50, or simply L. The data set used in this analysis was taken from a public database of over 25,000 nonconfidential compounds located at [ftp://epnws1.ncicfcrf.gov/gi50\\_m97.bin](ftp://epnws1.ncicfcrf.gov/gi50_m97.bin).

### 1. A Simplified Model

The basic hypothesis from molecular biology that accounts for the data is that the foreign compound finds its way into a cell and selectively binds with a protein enzyme, effectively inhibiting a critical metabolic pathway and thereby inhibiting the growth of the cell. To attempt a quantitative prediction, a detailed kinetic model could be developed, at least for the critical reaction step. One could then hypothesize that a given reaction step is related to the cell growth rate linearly. The end of this process could yield the predicted value of the I50 depending on many unknowns in a particular cell and for a particular mode of action including the reaction rate constants, concentrations of substrate, enzyme, and product, and reaction velocities of the neighboring reaction steps. Theoretically, assuming that the basic hypothesis is correct, it would then be possible to fit this monstrous equation to the observed data to find least squares estimates for the unknown parameters.

Part of the program to fit the data involves categorizing compounds by like-mode of action. This is necessary because the interaction of the compound with the enzyme is the same in all cells. On the other hand, the sensitivity to a particular mode of action varies from cell to cell. This suggests an initial analysis that bypasses a rigorous derivation for the I50 and all of the unknowns involved. We therefore postulate that to each compound  $m$  there is a particular mode of action  $p$ . We now hypothesize two measurable constants:  $a_{i,p}$  measures the sensitivity of cell  $i$  to the mode of action  $p$ , and  $b_{m,p}$  measures the reactivity of compound  $m$  with its mode of action. A given cell  $i$  will have many modes of action and many sensitivities  $a_{i,p}$ , while a given compound will have only one mode of action  $p$  and therefore only one reactivity constant  $b_{m,p}$ . We do not specify the units or scale for these constants, except that they will range from 0 to  $\infty$  and have an expected interpretation. The purpose is that we now hypothesize that the observed I50 of compound  $m$  acting on cell  $i$  is given by

$$I50_{i,m} = \frac{1}{a_{i,p}b_{m,p}}$$

where  $p$  is the mode of action of  $m$ . The law that is being proposed is that the I50 of a compound is inversely proportional to the reactivity of the compound and inversely proportional to the sensitivity of the cell to the mode of action.  $L_{i,m}$  will denote the logarithm of  $I50_{i,m}$  plus a normally distributed random variable  $\epsilon_{i,m}$  with mean 0 and standard deviation  $\sigma$  modelling the experimental error of measurement.

Now consider two compounds  $m$  and  $n$  acting on a cell  $i$ . If the compounds both have the same mode of action then

$$\begin{aligned} L_{i,m} - L_{i,n} &= (-\log a_{i,p} - \log b_{m,p} + \epsilon_{i,m}) - (-\log a_{i,p} - \log b_{n,p} + \epsilon_{i,n}) \\ &= (\log b_{n,p} - \log b_{m,p}) + (\epsilon_{i,m} - \epsilon_{i,n}). \end{aligned}$$

The point here is that the difference does not depend on the cell. For this reason, would expect that the observed value of  $L_{i,m} - L_{i,n}$  should in fact be distributed normally with mean  $(\log b_{n,p} - \log b_{m,p})$  and variance  $2\sigma^2$ . Consequently, if the variance of this observed value is taken over all cells, we should see a value close to  $2\sigma^2$ . On the other hand, if  $m$  and  $n$  act on cell  $i$  and have different modes of action  $p$  and  $q$  respectively, then

$$\begin{aligned} L_{i,m} - L_{i,n} &= (-\log a_{i,p} - \log b_{m,p} + \epsilon_{i,m}) - (-\log a_{i,q} - \log b_{n,q} + \epsilon_{i,n}) \\ &= (\log b_{n,p} - \log b_{m,q}) + (\log a_{i,q} - \log a_{i,p}) + (\epsilon_{i,m} - \epsilon_{i,n}). \end{aligned}$$

This does depend on the cell, assuming  $a_{i,q}$  and  $a_{i,p}$  are not correlated (the modes of action are distinct). Consequently, the variance of this observed value over all of the cells should be very high.

The statistical hypothesis in its most general form is that the data comes from a random variable  $L_{i,m} = -\log a_{i,p} - \log b_{m,p} + \epsilon_{i,m}$  where the  $a_{i,p}$  and  $b_{m,p}$  are fixed numbers depending on  $i, m$  and an unspecified mode of action  $p$ , and  $\epsilon_{i,m}$  is a normal variate with mean 0 and standard deviation  $\sigma$ . We propose to test this hypothesis by observing the statistic

$$\Delta = \frac{1}{N} \sum_{i=1}^N \left( \delta_i - \frac{1}{N} \sum_{j=1}^N \delta_j \right)^2$$

where  $\delta_i = L_{i,m} - L_{i,n}$ . (This statistic will be denoted  $\Delta$  when the compounds are assumed to have the same mode of action and  $\Delta'$  when they do not.) As we have already suggested, we expect this statistic to have a bimodal distribution under the hypothesis, and the remainder of this paper will continue with a more rigorous derivation of the power of this test.

## 2. Statistics for Like-Mode Compounds

For like-mode compounds, we have from the definition

$$\delta_i = (\log b_{n,p} - \log b_{m,p}) + (\epsilon_{i,m} - \epsilon_{i,n}).$$

Then from the expression for  $\Delta$  it is plain to see that

$$\delta_i - \frac{1}{N} \sum_{j=1}^N \delta_j = \epsilon_{i,m} - \epsilon_{i,n} - \frac{1}{N} \sum_{j=1}^N \epsilon_{j,m} - \epsilon_{j,n}.$$

Now we will assume that a particular  $m$  has been chosen and we want to know about the distribution of  $\Delta$  as it varies over other like-mode compounds in the sample. To that end we define the random variable

$$\epsilon_i = \epsilon_{i,m} - \epsilon_{i,n}$$

where  $\epsilon_{i,m}$  is assumed to be a fixed sampled constant and  $\epsilon_{i,n}$  is allowed to be a normally distributed random variable with standard deviation  $\sigma$ . Therefore  $\epsilon_i$  is normally distributed with mean  $\epsilon_{i,m}$  and standard deviation  $\sigma$ . The discriminating statistic becomes

$$\Delta = \frac{1}{N} \sum_{i=1}^N \left( \epsilon_i - \frac{1}{N} \sum_{j=1}^N \epsilon_j \right)^2.$$

The distribution of  $Q = N\Delta/\sigma^2$  is commonly known to be a noncentral  $\chi^2$  with  $N - 1$  degrees of freedom and noncentrality parameter  $Q_0$  (the value of  $Q$  obtained by setting  $\epsilon_{i,n} = 0$ ). We write  $Q_0 = N\Delta_0/\sigma^2$ . For reference purposes, if  $Q$  is  $\chi^2(r, \lambda)$  then  $EQ = r + \lambda$  and  $VarQ = 2r + 4\lambda$  (for an elementary derivation, see the derivation in the first section).

We conclude that

$$\begin{aligned} E\Delta &= \frac{N-1}{N} \sigma^2 + \Delta_0 \\ Var\Delta &= 2 \frac{N-1}{N^2} \sigma^4 + \frac{4}{N} \sigma^2 \Delta_0 \end{aligned}$$

Now consider the form that  $E\Delta$  and  $Var\Delta$  take when  $N$ , the number of cells commonly tested, is sufficiently large. We have the limits

$$\begin{aligned} E\Delta &\rightarrow \sigma^2 + \Delta_0 \\ Var\Delta &\rightarrow \frac{4}{N} \sigma^2 \Delta_0 \end{aligned}$$

These two equations show that the distribution for  $\Delta$  depends on the systematic measurement error  $\sigma$  and the particular error of the comparison compound given by  $\Delta_0$ . It is unlikely that any comparison

compound would be without error, and it is very possible for a particular compound to have unexpected error with large  $\Delta_0$ , and the effect is also shown in these equations. In particular, unmodelled errors will increase both the mean and variance of  $\Delta$ . The effect of increasing the mean is to push  $\Delta$  from like-mode molecules into the region of unlike-mode compounds making it difficult to distinguish. At the same time, as  $Var\Delta$  increases, the effect is to flatten out the  $\Delta$  histogram for like-mode compounds again making it difficult to ascertain a division between the unlike-mode compounds.

Assuming that unmodelled errors are rare, we can predict the effect on  $\Delta$  of the typical comparison compound. First note that  $N\Delta_0/\sigma^2$  has a classic  $\chi^2(N-1)$  distribution. So the observed  $\Delta_0$  will have

$$E\Delta_0 = \frac{N-1}{N}\sigma^2$$

$$Var\Delta_0 = 2\frac{N-1}{N^2}\sigma^4.$$

Therefore the typical comparison compound will have  $\Delta$  values that spike at  $2\sigma^2$  for all like-mode compounds.

### 3. Statistics for Unlike-mode Compounds I

We now consider the case where the comparison compound  $m$  is compared to unlike-mode compounds which will be denoted by the symbol  $n$ . According to the general hypothesis, there are an unspecified number of modes of action different from the comparison molecule to which  $n$  can belong. The actual distribution of  $\Delta'$  therefore depends on their aggregate. But too much is unknown about the remaining modes of action to hope to be able to infer anything from the observed distribution of  $\Delta'$ . Therefore we will modify the statistical hypothesis and assume that the values of  $L_{i,n}$  are distributed independently of  $L_{i,m}$ . It is therefore most natural to hypothesize that the variables  $L_{i,n}$  for  $i = 1, \dots, N$  are normal multi-variates.

To begin with a simplified derivation, suppose that  $L_{i,n}$  is normally distributed with mean  $\mu_i$  and common standard deviation  $\tau$  or  $n(\mu_i, \tau)$ . Then the random variable  $\delta_i = L_{i,m} - L_{i,n}$  is distributed as  $n(L_{i,m} - \mu_i, \tau)$ . With

$$\Delta' = \frac{1}{N} \sum_i \delta_i^2 - \frac{1}{N^2} (\sum_i \delta_i)^2$$

the distribution of  $Q' = N\Delta'/\tau^2$  is  $\chi^2(N-1, Q'_0)$ . So letting

$$\Delta'_0 = \frac{1}{N} \sum_i (L_{i,m} - \mu_i)^2 - \frac{1}{N^2} (\sum_i L_{i,m} - \mu_i)^2$$

We have

$$E\Delta' = \frac{N-1}{N}\tau^2 + \Delta'_0$$

$$Var\Delta' = 2\frac{N-1}{N^2}\tau^4 + \frac{4}{N}\tau^2\Delta'_0.$$

Similarly, we have limiting behavior of

$$E\Delta' \rightarrow \tau^2 + \Delta'_0$$

$$Var\Delta' \rightarrow \frac{4}{N}\tau^2\Delta'_0.$$

Again considering the choice of comparison compound as variable, the distribution of  $N\Delta'_0/\tau^2$  is  $\chi^2(N-1)$  so

$$E\Delta'_0 = \frac{N-1}{N}\tau^2$$

$$Var\Delta'_0 = 2\frac{N-1}{N^2}\tau^4.$$

Therefore the typical comparison compound will have  $\Delta'$  values that spike at  $2\tau^2$  for all unlike-mode compounds.

Two things are working to make the predicted distribution of  $\Delta'$  much different than  $\Delta$ . Both the expected value and the variance are made much larger than that of  $\Delta$  because

1.  $\tau$  measures the variance of the L50s of the population, and can be expected to be much larger than  $\sigma$  which is the variance of the experimental error for a single measurement.
2.  $\Delta'_0$  measures the deviation of the L50s of the comparison compound from the entire sample population. Therefore it should be expected to be much larger than  $\Delta_0$  which measures the deviation of the observed L50s of the comparison compound from its true values.

The expectations and variances of  $\Delta$  and  $\Delta'$  show that their distributions will ideally be located in different ranges of values. This should make it possible to distinguish the two types of compounds based on the value of  $\Delta$  and the area of the distribution in which it falls.

#### 4. Statistics for Unlike-mode Compounds II

Unfortunately, the prediction for unlike-mode compounds assuming a common standard deviation are not even close to the observed data. The value of  $\tau$  that is computed from the database is reasonable (all L50 values are approximately equal to  $-4$  with a variance of about 1), but it predicts a mean for the unlike-mode distribution that is several orders of magnitude too large. This immediately suggests that the assumption of independence and common variance was much too simplified. Indeed, a quick look at the covariance matrix shows that all cells are very strongly correlated. Therefore, a better analysis will take the covariance matrix into account in predicting the distribution of unlike-mode compounds.

Now assume that  $\mathbf{L}_n = (L_{1,n}, \dots, L_{N,n})$  is a multivariate normal distribution with means  $\mu = (\mu_1, \dots, \mu_N)$  and covariance matrix  $\mathbf{V}$ . Note that  $\Delta$  is a quadratic form on  $L_{i,m} - L_{i,n}$  which has a multivariate normal distribution with means  $\mathbf{L}_m - \mu$  and covariance  $\mathbf{V}$ . The resulting quadratic form does not necessarily have a non-central  $\chi^2$  distribution (although the general shape of the distribution will be similar). Therefore we cannot appeal to the formulas for  $E\Delta$  and  $Var\Delta$ , but it is possible to derive formulas directly (see the derivation in the second section). These formulas require the matrix defining the quadratic form,

$$\mathbf{A} = \frac{1}{N}(\mathbf{I} - \frac{1}{N}\mathbf{1})$$

where  $\mathbf{1}$  is the  $N \times N$  matrix whose entries are all 1. Note that  $\mathbf{A}^2 = 1/N\mathbf{A}$  and since  $\mathbf{VA}$  is symmetric,  $\mathbf{AVA} = 1/N\mathbf{VA}$ . Using the formula derived,

$$E\Delta = tr(\mathbf{VA}) + (\mathbf{L}_m - \mu)' \mathbf{A} (\mathbf{L}_m - \mu)$$

$$Var\Delta = 2tr(\mathbf{VA})^2 + \frac{4}{N}(\mathbf{L}_m - \mu)' \mathbf{VA} (\mathbf{L}_m - \mu).$$

Interpreting these formulas is a little more difficult than in the first analysis. The terms involving  $tr(\mathbf{VA})$  and  $tr(\mathbf{VA})^2$  here are very similar to the corresponding terms in the noncentral  $\chi^2$  formula, except that  $\tau^2$  is replaced with a value that is reduced by cell-cell correlation. The second term in  $E\Delta$  is simply  $\Delta'_0$ . It can also be seen that the second term of  $Var\Delta$  is much like  $4/N\tau^2\Delta'_0$  but with  $\tau^2$  reduced by dependence.

We now have a less certain prediction that the distributions of like-mode and unlike-mode  $\Delta$  values will be significantly different, but it is nevertheless a prediction that can be tested. This involves computing the means  $\mu$  and the covariance matrix  $\mathbf{V}$ . Technically, the mean and covariance in the derivation was assumed to be valid for the unlike-mode compounds only. There is no way to estimate them, however, without including the like-mode compounds, and this may bias the result. Then for a given comparison molecule, the mean and variance of the  $\Delta'$  statistic is computed and these parameters are compared with the observed histogram of  $\Delta$  values. In particular, using the method of moments, a  $\Gamma$ -distribution can be superimposed on the empirical distribution to see if there is a close fit with one part.

#### 5. Statistics for Unlike-mode Compounds III

Once again the predicted mean and standard deviation are not sufficiently close to the observed distribution of  $\Delta$  (bimodal or otherwise). Apparently there is a complication that we have not considered yet due to incomplete data. In particular, the data set does not include a value for every compound and every cell: some are missing. The calculation of  $\Delta$  as well as the cell means and covariance matrix, and predicted means and standard deviations from the previous analysis can still be performed, but the prediction loses its meaning.

To take missing data into consideration, let  $s$  denote a subset of integers  $\{1, \dots, N\}$  and let  $\Delta_s$  denote the variance statistic

$$\Delta_s = \frac{1}{|s|} \sum_{i \in s} \delta_i^2 - \frac{1}{|s|^2} (\sum_{i \in s} \delta_i)^2$$

where for the moment we assume that the entire population is from unlike-mode compounds and we drop the prime from  $\Delta$ . The statistic being observed is  $\Delta = \Delta_S$  where  $S$  is a random variable whose values are subsets  $s$ . The analysis of the previous section applies to each  $\Delta_s$ , and we denote its mean by  $E\Delta_s$  and its variance by  $Var\Delta_s$ . From the general theory of sampled statistics (see the derivation in the third section ) we know that  $\Delta$  has mean

$$E\Delta = E_S(E\Delta_s)$$

$$Var\Delta = E_S(Var\Delta_s) + Var_S(E\Delta_s)$$

where the subscripts on  $E_S$  and  $Var_S$  signify that the mean and variance are taken over the random variable  $S$ .

At this point we forego the explicit calculation. Although it is possible to obtain a simple statistic for the  $E_S$  terms, we have little hope of making qualitative predictions. We must content ourselves with numerical estimates of the predicted distribution and the resulting goodness of fit. In fact, we are able to gain much more information about the predicted distribution by abandoning the analytic approach all together. In its place we use a Monte Carlo simulation to generate values of  $\Delta$  based on a multivariate normal distribution of  $L_{i,m}$  for only those non-missing values in the data set. That completely describes how we propose to generate the predicted distribution and so we can finally turn to testing the prediction.

The same analysis of missing data applies to the variance statistic for like-mode compounds.

## 6. Estimating the Covariance

To summarize the problem in abstract terms, we hypothesize that the data in the form  $\delta$  is sampled from a distribution that is a mixture of unknown proportions of two different multi-variate normal distributions with unknown means and covariance matrices. And if that were not bad enough, the sampled data is incomplete, i.e. each sample has components whose values are not given. Undoubtedly, there is a systematic reason that components have not been measured or supplied, although for the time being we will assume that components have been suppressed by an independent random process.

To this point, we have developed increasingly complicated models in an effort to fit the given data and allow a test of the hypothesis. In its current state, we require a Monte-Carlo simulation of a multi-variate normal distribution, which in turn requires a proposed mean and covariance matrix. To generate multi-variate normal deviates, we require the Cholesky decomposition of  $\mathbf{V}$ , namely  $\mathbf{V} = \mathbf{L}\mathbf{L}'$  where  $\mathbf{L}$  is a lower triangular matrix. This is possible, of course, because  $\mathbf{V}$  is hypothetically symmetric and positive-definite.

The usual expression for estimating  $\mathbf{V}$  from a sample with complete data produces an unbiased estimate in the form of a symmetric positive-definite matrix. The estimate of  $\mathbf{V}$  based on incomplete data is another story all together. What is certain, however, is that estimating  $\mathbf{V}$  by simply counting only the available data does not produce a positive-definite matrix. And without a positive-definite matrix, we are unable to generate multi-variate normal deviates.

## 7. Testing the Prediction

Several parameters are implicit in the model we have developed. To begin with, we can determine the sample means, and apply what is known about incomplete data to estimate a symmetric positive-definite covariance matrix. With the Cholesky decomposition, and a generator for normal deviates, we can create a Monte-Carlo simulation (of given data only) to predict the distribution of  $\Delta$  for unlike-mode compounds. We must then estimate the proportion of compounds in the dataset having unlike-mode of action. We do this by postulating that *under favorable conditions*, fewer than half of all compounds in the dataset are like-mode compounds, and the corresponding value of  $\Delta$  is below the population median. Therefore the upper half of the observed distribution of  $\Delta$  should be very close to a multiple of the predicted distribution in the same range. The multiple is precisely the number of predicted unlike-mode compounds and the goodness of fit measure (e.g. a  $\chi^2$ ) gives a first test of our hypothesis. Assuming that there is a good fit, the predicted distribution of  $\Delta'$  can be subtracted from the observed distribution of  $\Delta$ , and what is left should fit a non-central  $\chi^2$ . Missing data enters into the prediction for this distribution and it poses a problem for estimating the parameters and finding a goodness of fit. We require at the very least that the resulting distribution of like-mode compounds is sharp compared to the distribution of unlike-mode compounds. Therefore we will compute the variance of the like-mode distribution and compare it with the variance of the unlike-mode distribution. In fact, by using the method of moments, we can compute values for  $\sigma$  and  $\Delta_0$  as if the like-mode compounds were actually distributed as a non-central  $\chi^2$  (which may be valid if the variation

introduced by missing data is small), and then test this hypothesis directly. A confirmation would surely imply a confirmation of the general hypothesis.

However the goodness of fit is tested, it is not necessary to identify any parameters to deduce a probability of membership in the class of like-mode compounds. After all, for a given value of  $\Delta$  we know the number that are predicted to belong to the class of unlike-mode compounds and the remaining proportion may be taken to belong to the like mode compounds.

### 8. Things To Do

0. Learn how to compute a symmetric positive-definite matrix estimate for the covariance matrix.
1. Use a Monte-Carlo simulation to compute a predicted unlike-mode distribution.
2. Estimate the number of compounds in the unlike-mode branch of the distribution by examining the upper median of the data.
3. Determine the goodness of fit to the upper median of the distribution.
4. Subtract the unlike-mode distribution from the empirical distribution and fit the result to a noncentral  $\chi^2$  (approximate). Test this goodness-of fit.
5. For a good fit, determine the probability that a compound has the same mode of action as the comparison compound based on the value of  $\Delta$ . *I hope to complete the steps to this point for a seminar on June 29.*
6. Generate synthetic data and test the method for accuracy and sensitivity to the choice of comparison compound.
7. Determine by analysis and experiment how the inclusion of like-mode compounds in the estimates for  $\mu$  and  $\mathbf{V}$  bias the result. If significant, design a method to iteratively correct the estimates and remove the bias.
8. In the end, most compounds should be assigned to one of a few different proposed major modes of action, while some compounds may have a unique mode of action. Compare categorization as the comparison molecule is varied.
9. For each major mode of action  $p$ , use the corresponding subset of data to determine a least squares estimate for  $a_{i,p}$  and  $b_{m,p}$  for each cell and each compound having that mode of action.
10. Compare the estimated values of  $a_{i,p}$  for each cell with other measures of cell metabolism, e.g. target data, to see if there is a correlation with any in particular. A strong correlation with a single target indicates that the mode of action for those compounds implicates that target.
11. Look for other kinds of quadratic forms (e.g. weighted) that *accentuate* the differences between the like-mode and unlike-mode compounds. For example, correlated cells should probably be grouped and averaged so that they don't overshadow the effect of unique under-represented cells.