A New Method for the Statistical Analysis of Dual Labeled Isotope Polypeptides Separated by Gel Electrophoresis

Sanford Weisberg*

Technical Report No. 202

March 1973

*The author is at the Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55101. This work was supported in part by N.S.F. Grant GS-32327x to the Statistics Department, Harvard University. The data used come from research in The Biological Laboratories, Harvard University, under the direction of R.P. Levine. The work is supported by grants to R.P. Levine from the N.S.F. (GB-29203) and the Maria Moors Cabot Foundation for Botanical Research at Harvard University. The author would also like to thank William G. Burton and David Hoaglin.

I. SUMMARY

Analysis of cellular polypeptides following separation by electrophoresis on polyacrylamide gels (7) often takes advantage of radioactive isotopes that are introduced into the polypeptides during cell growth. Comparisons of polypeptides between differing organisms, for example, wild type versus mutant (8), or studies of the rate of turnover of polypeptides (1), conveniently involve the use of two radioactive labels (such as ^{3}H and ^{14}C). The gels on which the polypeptides have been separated are cut into 1 mm slices and the radioactivity associated with each slice is determined and expressed as ^{3}H or ^{14}C counts per minute (cpm). This paper described a simple method of detecting differences in the polypeptides in the two differently labelled organisms.

II. MOTIVATION

After electrophoresis, let ^{14}C and ^{3}H denote the cpm in each gel slice, after correction for background and crossover (4). First consider an experiment in which both ^{3}H and ^{14}C are used to label polypeptides from identical organisms. We shall call this a control experiment. In control experiments, we expect the ratio of ^{3}H to ^{14}C cpm in each slice to be a constant, which we will call r. The parameter r is essentially a function of the amount of radioactive material placed on the gel. Thus, for control experiments, we expect the following equation to hold in every slice

$$(1) \qquad \frac{^{3}H}{^{14}C} = r.$$

In experiments in which ^{3}H and ^{14}C label polypeptides from different organisms (such as a mutant and a wild type respectively), the mathematical relationship (1) will fail to hold in some slices, indicating those polypeptides by which the two organisms differ.

Even in a control experiment, we will not observe exactly r for the ratio in each slice. Neither the measurement process nor the experimental method are without error, so that the value of r in each slice will vary. Furthermore, the counts for each polypeptide will be distributed over several slices; thus, even if there were no experimental error, we would still find variability in the ratio $^{3}H/^{14}C$.

In principle, we could take equation (1), estimate the value of r, and assess the fit of the model at each slice. This approach is very difficult because ratios of the form $^{3}H/^{14}C$ are very unstable, ranging from 0 to ∞. Furthermore, the variability of the ratio is highly

dependent upon the size of each of the counts. Finally, no standard, widely accepted methods exist for handling these ratios.

A more useful approach can proceed as follows: We multiply both sides of equation (1) by ^{14}C to get the equation

$$(2) \quad {}^{3}H = r\ {}^{14}C$$

We immediately recognize the following fact: In a control experiment, in each slice, the ^{3}H cpm should be a constant multiple of the ^{14}C cpm. Equivalently, if we were to graph the ^{14}C cpm on the X axis against the ^{3}H cpm on the Y axis, we would expect the graph to be a straight line with slope equal to r. As an example, in one control experiment where both ^{3}H and ^{14}C label the same organism, this graph appears as Figure 1.

[Figure 1 about here]

Notice several important characteristics of Figure 1. First, the graph of the points is basically a straight line, as predicted by equation (2). If we were to fit a line to the points by eye, about as many points fall above the line as below it. Finally, and most importantly, as the counts get larger, the variability between points also gets larger.

If a monotonic transformation of the data to a different scale is performed, it may be possible to make the variability of the data nearly independent of magnitude. For example, suppose the counts at each slice follow a Poisson distribution; that is, for ^{14}C in each slice,

$$P(^{14}C = k) = e^{-m}\ m^{k}/k! \qquad k = 0, 1, \ldots$$

where m = the expected ^{14}C count for that slice. (A similar law will hold for ^{3}H). This is the usual model for counted data. In this case

we will also have that the standard deviation of ^{14}C is equal to $\sqrt{m}$. Thus the variability of a count will increase as the square root of its size.

It can be easily shown [Scheffé, (6)] that, if Y follows a Poisson distribution with mean m (and therefore standard deviation $\sqrt{m}$), then the variable $\sqrt{Y}$ is distributed approximately normally with mean $\sqrt{m}$ and standard deviation about $\frac{1}{2}$. The important point for the purpose of this analysis is that, if the Poisson model is true,

then the graph of $\sqrt{^3H}$ vs $\sqrt{^{14}C}$ will be a straight line with the added benefit that the variability away from the line will remain constant as the size of the count increases. This graph for the same control experiment is shown as Figure 2.

[Figure 2 about here]

To summarize, in a control experiment, we expect equation (2) to hold for each slice number. To make variability <u>independent</u> of size of count, we take the square root of equation (2) and, letting $a = \sqrt{r}$, we get

$$(3) \quad \sqrt{^3H} = a\sqrt{^{14}C}$$

All further analysis will use equation (3).

III. ANALYSIS

In equation (3) we recognize a standard linear regression problem, which is to fit a straight line of the form

$$(4) \quad Y = aX + b$$

to a set of pairs of points (X,Y). However, the problem at hand differs from the standard problem in several ways. Usual regression problems are primarily interested in estimation of the parameters a and b. In our problem, a is simply an artifact of the experiment. Although it must be estimated, we consider it to be merely a nuisance parameter since it is a function of the amount of radioactive material on the gel.

The value of the parameter b in (4) is known to be zero since when $^{14}C = 0$ we would also have $^{3}H = 0$. However, we shall estimate b as if it were not zero because we are interested in the regression line in the region where the data points actually occur, and not in the fit of the line in the region of zero.

The principal interest in the analysis is to examine deviations of the observed (transformed) counts from the estimated line. If, for a set of contiguous slices, the deviations are "large" and of the same sign, we will be able to say that the model fails to fit in that region. Thus, in experiments where ^{14}C and ^{3}H label polypeptides from different organisms, regions of slices where large positive deviations occur will denote excesses of ^{3}H (and hence excesses of a specific polypeptide in the ^{3}H-labeled organism) and large negative deviations will indicate deficiencies of ^{3}H (and hence deficiencies of specific polypeptides in the ^{3}H-labeled organism).

IV. ESTIMATION OF PARAMETERS

The standard method for estimating the parameters a and b is by least squares. Denoting $Y = \sqrt{{}^{3}H}$, $X = \sqrt{{}^{14}C}$, our estimates of a and b are

$$(5) \qquad \hat{a} = \frac{\Sigma\ (X-\bar{X})(Y-\bar{Y})}{\Sigma\ (X-\bar{X})^2} \quad ; \quad \hat{b} = \bar{Y} - \hat{a}\bar{X}$$

where the summation is over all slices, and $\bar{X}$ and $\bar{Y}$ denote the arithmetic mean of the X's and Y's respectively. This method of estimation is probably not the "best" under the circumstances since heavy weight is given to the extreme observations that may occur in experiments where ${}^{3}H$ and ${}^{14}C$ label different organisms. Much current statistical literature, such as (2) and (3), is concerned with techniques that are "robust" even when extreme observations occur. In the present problem, probably little additional information is to be gained by using a more resistant method of estimation compared to the loss due to increase in complexity.

We shall therefore estimate the regression line as

$$(6) \qquad \sqrt{{}^{3}H} \;=\; \hat{a}\sqrt{{}^{14}C} + \hat{b}\ ,$$

with $\hat{a}$ and $\hat{b}$ defined by (5).

For each slice, we can create a <u>deviation score</u>, <u>D</u>, such that

D = (observed value of $\sqrt{{}^{3}H}$) minus ("fit" value of $\sqrt{{}^{3}H}$)

or, for each slice,

$$(7) \qquad D = \sqrt{{}^{3}H} - (\hat{a}\sqrt{{}^{14}C} + \hat{b})\ .$$

The remainder of the analysis uses the D's. We make the following observations:

a. $\Sigma\ D = 0$, where the sum is over all slices. This is an artifact of using least squares to estimate a and b.

b. The variability in the D's is independent or nearly independent of both slice number and magnitude of the count in that slice, as is shown in Figure 2. This means that a "large" value of D in slice 52 has exactly the same meaning as a large value of D in slice number 93 or any other slice.

c. An estimate of variability of the D's can be obtained as $s = \sqrt{\Sigma\ D^2/n(n-2)}$, where the summation is over all slices. The statistic s is usually called the standard error of regression. In control experiments, s should be a reliable estimate of the real standard deviation of the D's. In an experiment where ^{3}H and ^{14}C label different organisms, the computation of s will put too much weight on those values of D that are large (i.e. those places where the model (3) fails) and will therefore overestimate variability.

d. In a control experiment, each D is equally likely to be positive or negative. This implies that, in experiments where ^{3}H and ^{14}C label different organisms, long runs of D's of the same sign might indicate a difference in polypeptides between the two organisms.

e. Ignoring the fact that the D's are probably correlated, we would hope that the D's from a control experiment could be regarded as a sample of size n (= number of slices) from an approximately normal distribution, mean zero, standard deviation estimated by s. Thus we would expect to see no more than about 5% of the D's greater than 2s or less than -2s, and only perhaps 1% beyond 2.5s. These are only to serve as approximate rules, because of violations in the necessary assumptions. In experiments where ^{3}H and ^{14}C label different organisms, we will simply have too many large D's (i.e. more than 1% larger than 2.5s) and D's that are too large (i.e. some D's greater than 4s or so).

V. EXAMPLE: CONTROL EXPERIMENT

Consider a control experiment in which ^3H and ^{14}C label the same organism. For the data shown in Figure 2, the least squares fit of the regression line is

$$(8) \quad \sqrt{^3H} = 1.570\sqrt{^{14}C} + 0.758$$

with standard error s = 1.417. We next compute the D's. For example, in slice number 59, the ^3H count was 302 counts per minute while the ^{14}C count was 112. Thus, for slice 59,

$$\begin{aligned} D &= \sqrt{^3H} - (1.570\sqrt{^{14}C} + 0.758) \\ &= \sqrt{302} - (1.570\sqrt{112} + 0.758) \\ &= 17.370 - 17.347 \\ &= 0.033 \end{aligned}$$

which is a close fit in this slice.

Possibly the most informative single graph is shown in Figure 3. It gives the slice number on the X-axis and the deviation for that slice on the Y-axis. We note that the deviations oscillate apparently randomly between positive and negative values, as we would expect. The only possible exception to this would be in the region from about 26 to 32, where the deviations are all large and positive; indeed, the four largest positive D's occur at slices 28, 29, 30, and 31.

To get a stronger feel for the size of these deviations, Figure 4 gives a normal probability plot of the D's. If the spread of the D's were as expected (more formally, if the D's formed a sample of size n from a normal distribution) then the graph in Figure 4 should be approximately a straight line. As we see the four points at the high end of the curve are too far to the right to be thought of as lying on the line.

Thus, we are tentatively led to conclude that slices 28-31 are unusual, and leave interpretation to the biologist. Beyond this single observation, the graphs yield the expected picture of a control experiment.

[Figure 3 about here]

[Figure 4 about here]

VI. EXAMPLE: A COMPARISON OF THE CHLOROPLAST MEMBRANE POLYPEPTIDES OF A NORMAL AND A MUTANT ORGANISM

A wild-type strain of Chlamydomonas reinhardi labeled with ^{14}C is compared to a mutant strain called F-54 labeled with ^{3}H. The wild-type can perform chloroplast membrane-bound functions involved in ATP synthesis. The mutant strain has lost one of these functions and consequently it cannot synthesize ATP (5). It is of interest therefore to compare the composition of the membranes of the wild-type and the mutant organisms.

Figure 5 gives the data resulting after electrophoresis.

The counts shown here have been corrected for background and crossover.

The fit regression equation of $\sqrt{^{3}H}$ on $\sqrt{^{14}C}$ is

$$\sqrt{^{3}H} = 2.557\sqrt{^{14}C} - 2.293 ,$$

with standard error s = 3.335. As previously stated, s probably overestimates the variability because too much weight is given to large deviations.

[Figure 5 about here]

[Figure 6 about here]

Figure 6 is a graph of deviations against slice number (note that the scale of Figure 6 is not the same as Figure 3). Figure 6 does not seem to oscillate between positive and negative values as easily as does Figure 3. More importantly, there are several regions in which large deviations occur. In slices 52 to 55, the deviations are very large and negative (the deviation in slice 53 is about -14.4, or about -4s), indicating a deficiency of ^{3}H and hence a deficiency of specific polypeptide (s) in the mutant relative to the wild-type. Another less obvious region that will need careful consideration by the biologist is around slices 87 to 96.

[Figure 7 about here]

Finally, Figure 7 gives a normal probability plot of the deviations. Unlike Figure 4, the plot is more S-shaped than straight, with several points at both ends of the graph well away from lying on a line. This is the characteristic shape of normal plots when the deviations do not resemble a sample from a normal distribution. This will be the case in experiments that show differences between the ^{3}H-labeled organism and the ^{14}C-labeled organism.

VII. REFERENCES

1. Arias, I. M., D. Doyle and R. T. Schimke [1971]. Studies on the synthesis and degradation of proteins of the endoplasmic reticulum of rat liver. J. Biol Chem. 244, 3303.

2. Forsythe, A. B. [1972]. Robust estimation of straight line regression coefficients by minimizing p-th power deviations. Technometrics 14. 159.

3. Hoerl, A. E. and R. W. Kennard [1970]. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55.

4. Kobayashi, Y. and D. V. Maudsley, in Bransome, E. D. Jr. (ed.) [1970]. The Current Status of Liquid Scintillation Counting. New York: Grune and Stratton.

5. Sato, V. L., R. P. Levine and J. Neumann [1971]. Photosynthetic phosphorlyation in Chlamydomonas reinhardi: effects of a mutation altering an ATP synthesizing enzyme. Biochim. Biophys. ACTA 253, 437.

6. Scheffé, H. [1959]. The Analysis of Variance. New York: Wiley.

7. Shapiro, A. L., E. Vinuela and J. V. Maizel Jr. [1967]. Molecular weight estimation of polypetpide chains by electrophoresis in SDS-polyaclamide gels. Biochem. Biophys. Res. Commun. 28, 815-20.

8. Shapiro, B. M., Antonio G. Siccardi, Y. Hirota and J. François [1970]. On the process of cellular division in Escherichia coli. J. Mol. Biol. 52, 75.

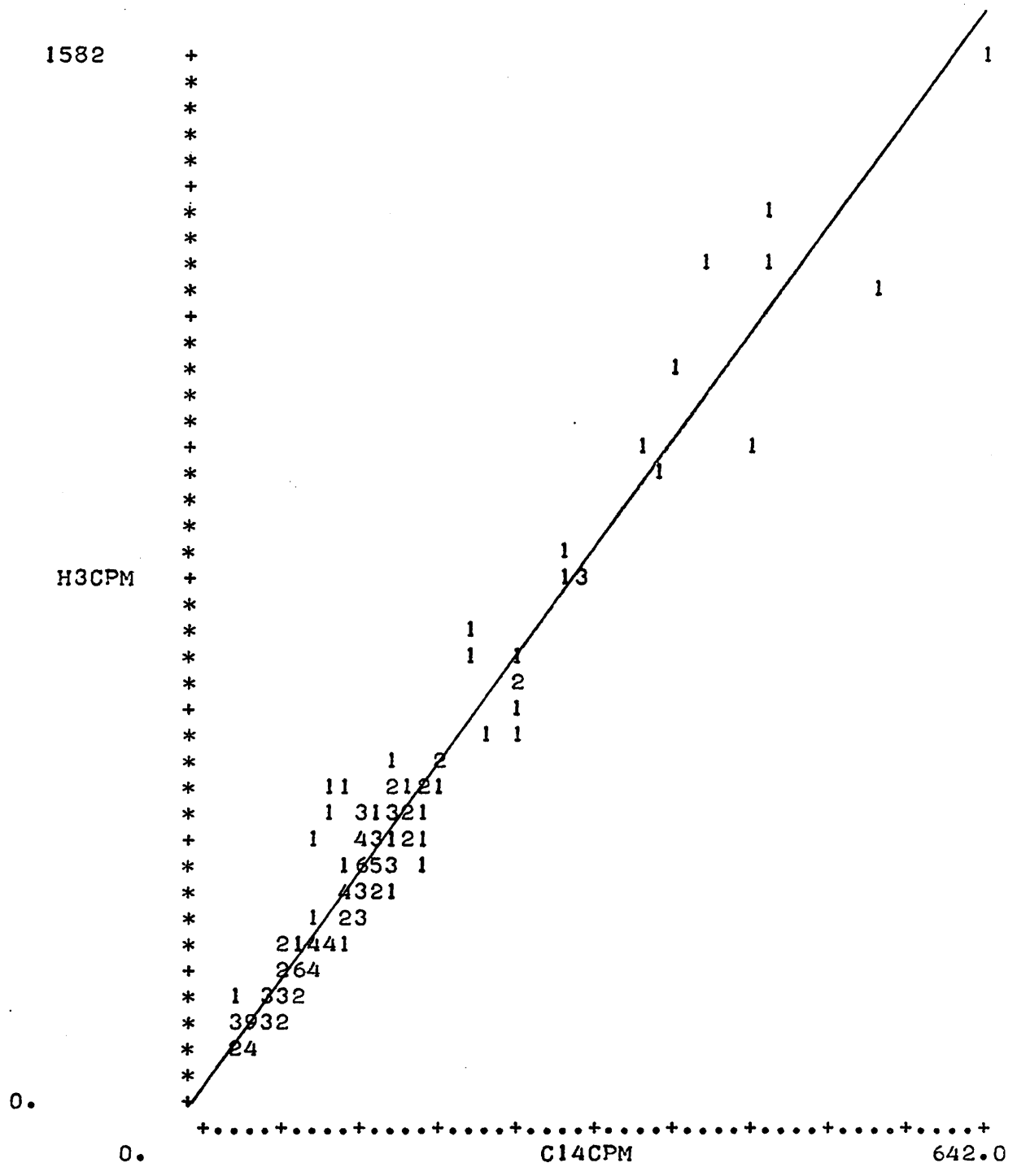


Figure 1. Graph of ^{14}C cpm (X-axis) versus ^{3}H cpm (Y-axis) for one control experiment. The fitted line is the least squares line. The numbers indicate the number of observations graphed at a point; e.g. a 4 indicates 4 points.

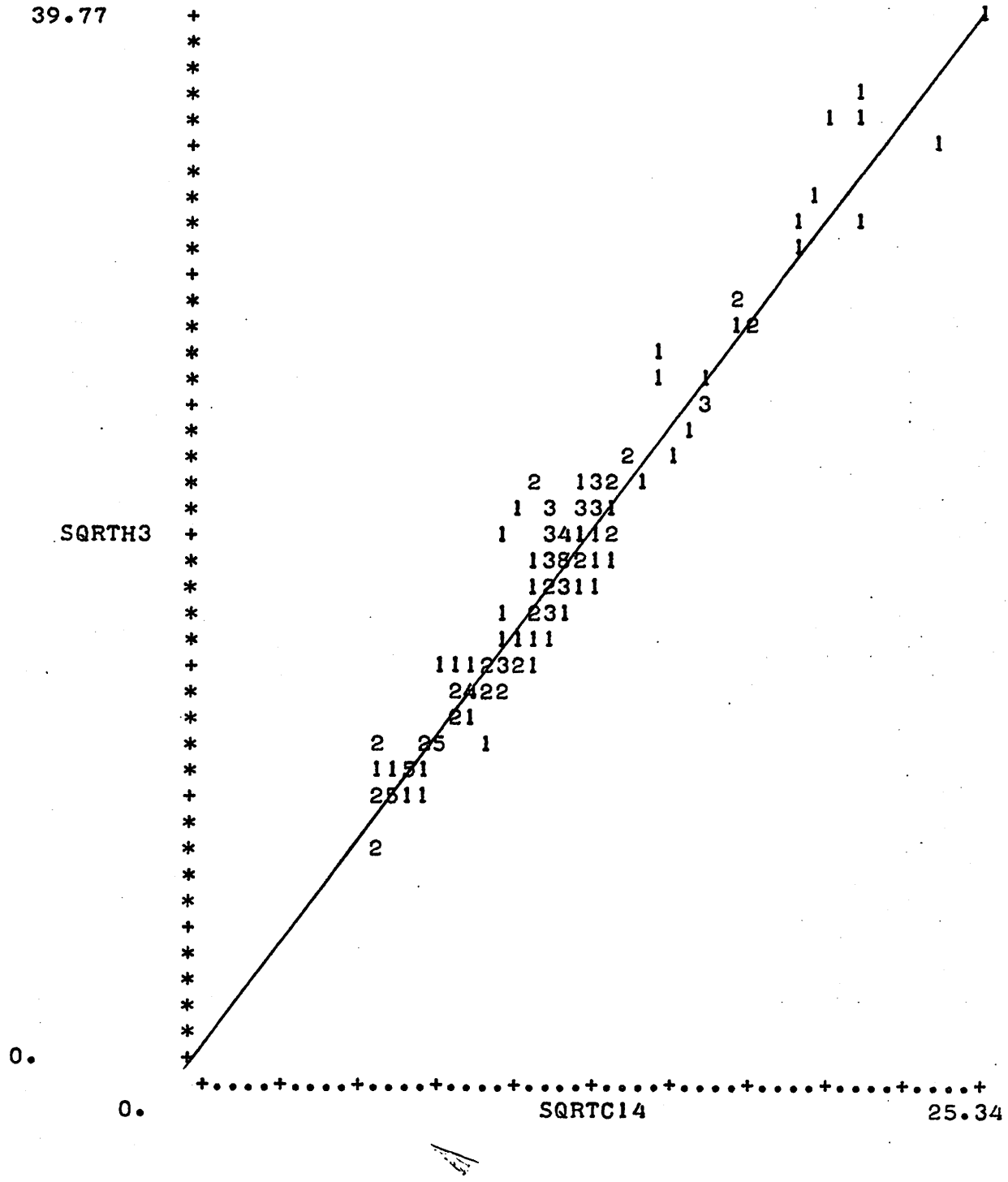


Figure 2. Graph of $\sqrt{^{14}C \text{ cpm}}$ (X-axis) versus $\sqrt{^{3}H \text{ cpm}}$ (Y-axis) for the same data as Figure 1. The fitted line is the least squares line.

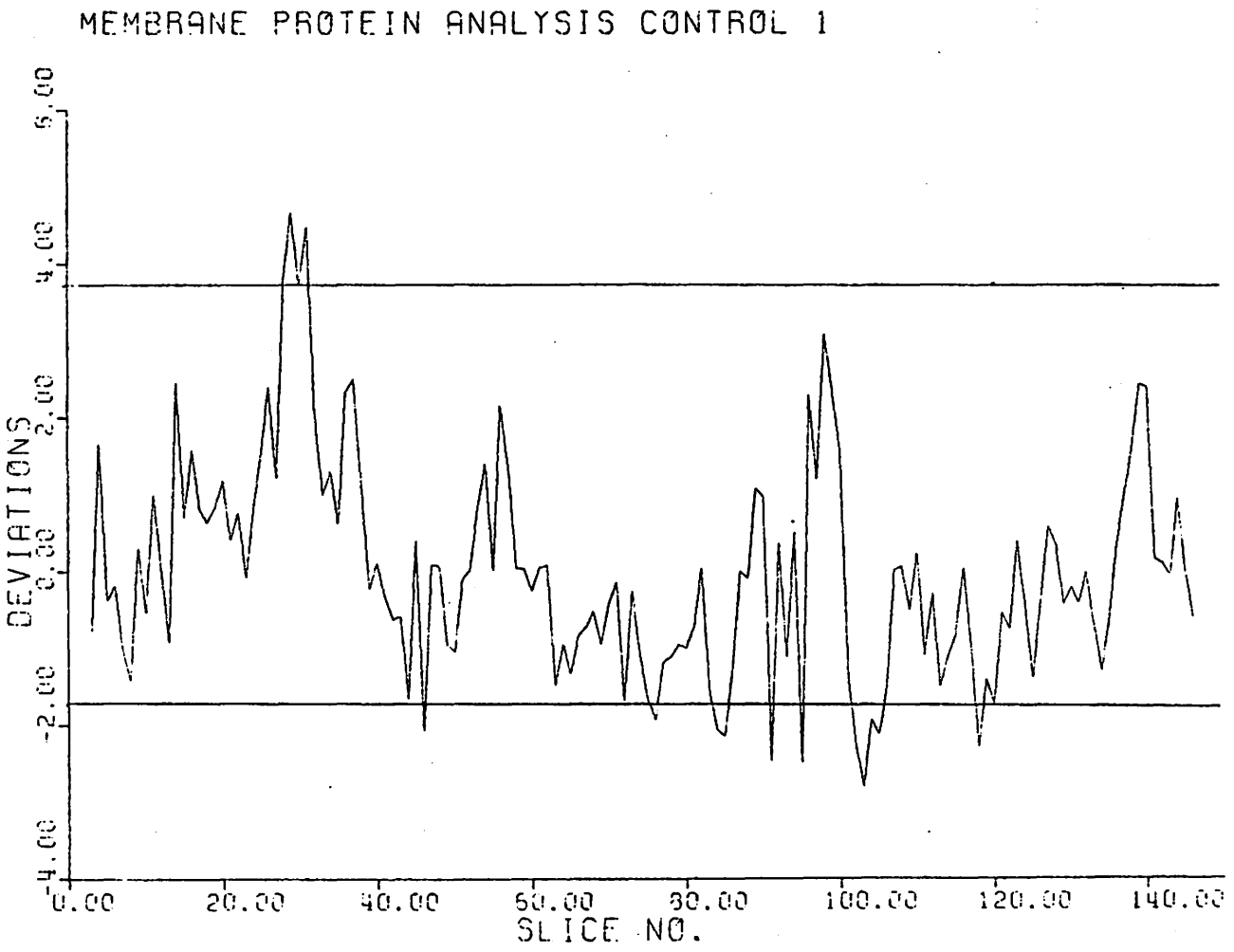


Figure 3. Graph of slice number (X-axis) versus deviation from regression in that slice (Y-axis) for a control experiment. The horizontal lines are drawn for convenience at $\pm$ 2s.

Figure 4. Normal plot of the deviations for a control experiment. The points graphed are $(D_{(i)}, \Phi^{-1}(\frac{i}{n+1}))$, where $D_{(i)}$ is the i-th smallest deviation, n is the number of slices, and Φ^{-1} is the inverse of the Standard Gaussian distribution function. The graph should approximate a straight line.

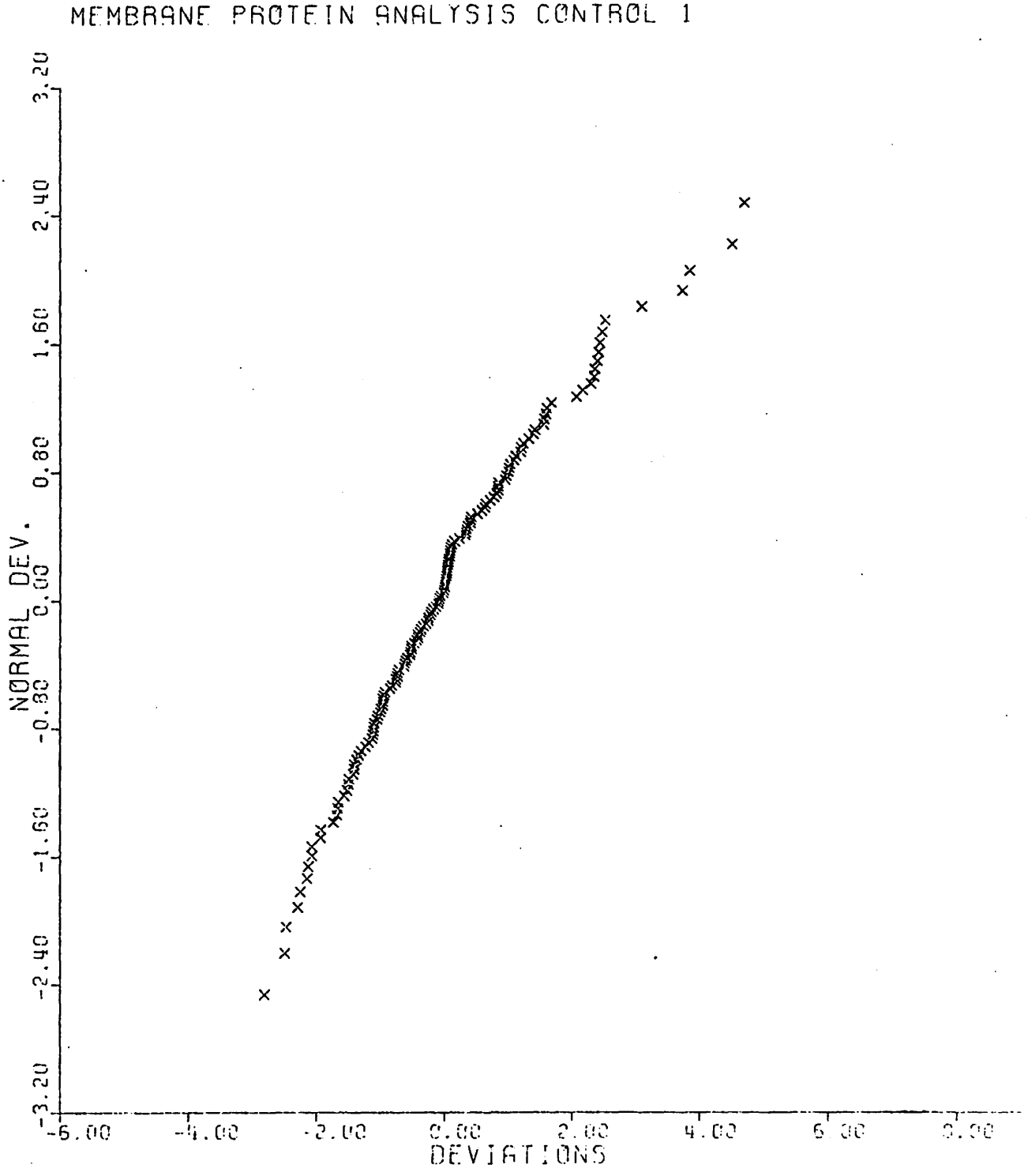
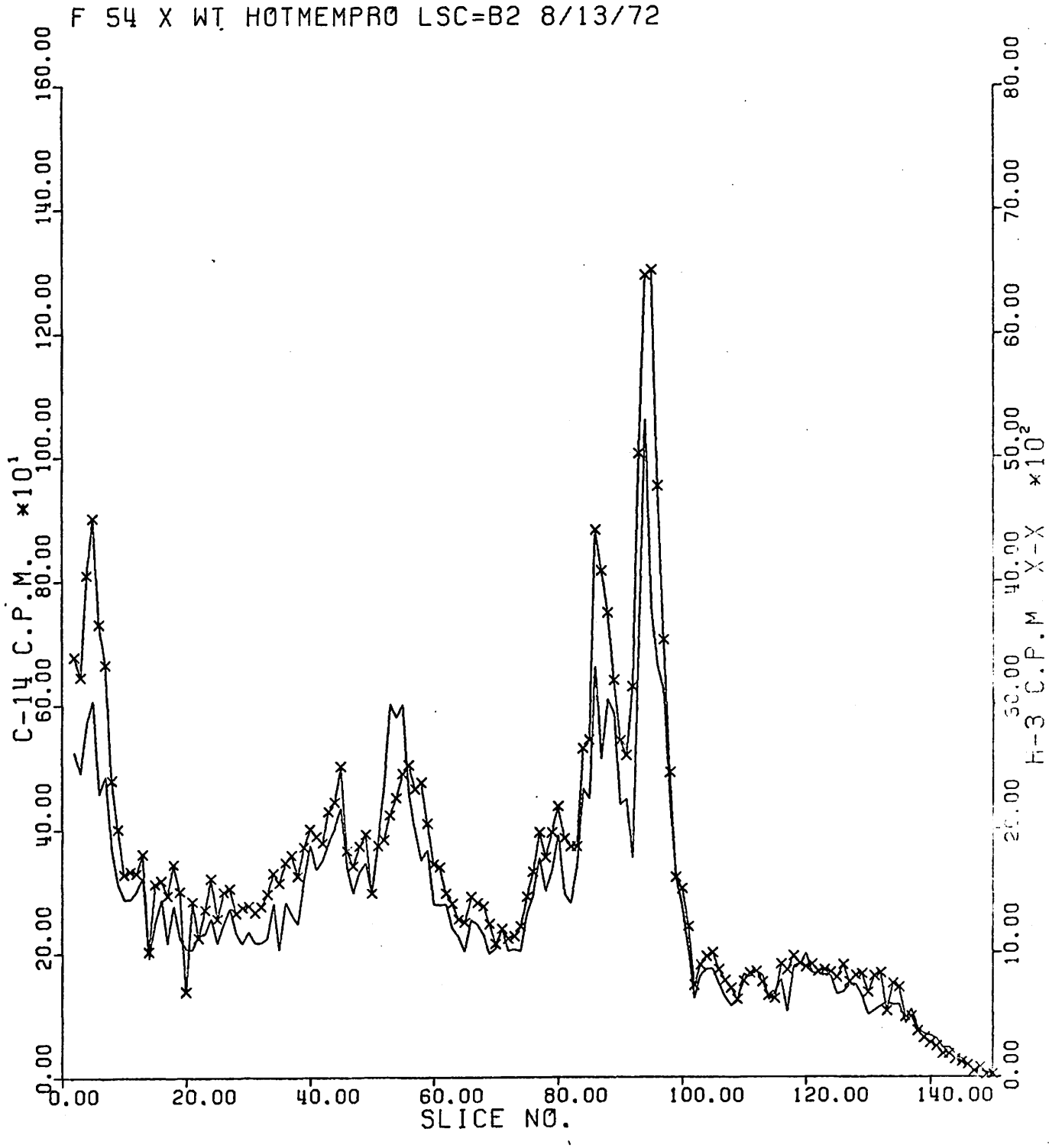


Figure 5. Plot of ^{14}C cpm (scale at left) and ^{3}H cpm (scale at right--these counts indicated by x's) for all slices used in the analysis (counts have been corrected for crossover and background).



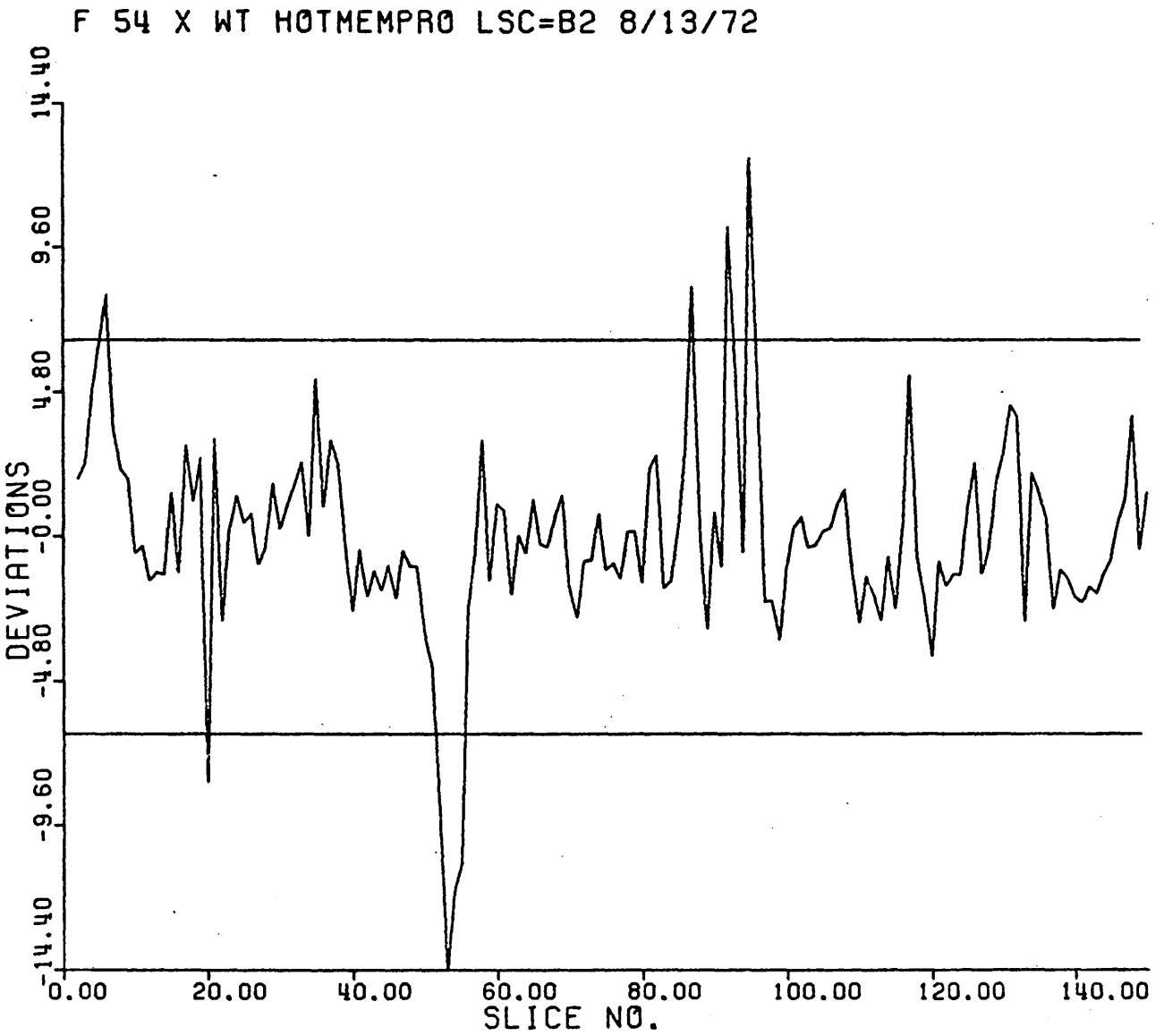


Figure 6. Slice number versus deviation from regression.

Figure 7. Normal plot.