Analysis of a 2^9 Full Factorial Chemical Library

S. Stanley Young* and Douglas M. Hawkins†

Information Technology, Glaxo Inc., Research Triangle Park, North Carolina 27709

Received March 17, 1995®

Robotic synthesis is making possible the synthesis of large, systematically designed sets of compounds. We analyze a 512-compound set that is a 2^9 full factorial experimental design using a recursive partitioning algorithm, FIRM, and a high-dimension visualization tool, TempleMVV. These techniques are used to quickly and easily identify the main trends in the data set and also identify unusual observations. We show that analytical and visualization methods can be used synergistically to analyze a large, complex, high-dimensional data set. We also show that a fractional factorial design of 128 compounds would give essentially the same information.

Introduction

We are at a pivotal point in the history of compound synthesis, and we have an opportunity to use chemical series design and analysis methods to obtain more complete information and understanding of important features of potential drugs. Robotic methods can be used to synthesize a large series of analogs at an economical cost. The chemical series can be planned on the basis of computational chemistry and statistical methods. The resulting systematic chemical series can be analyzed using powerful statistical methods to find general trends in the data set and isolate specific effects that may be the result of interactions among the molecular parts. But just as the opportunities are great, the problems are difficult. Which of the very many possible analogs should be made? What statistical experimental design methods are appropriate? How can compounds or parts of compounds be numerically characterized? We focus on the question of how large, complex data sets can be summarized and analyzed to extract important information. We do this through the analysis of an example.

Substance P is a neuropeptide that has been studied since 1970. A complete data set of 512 analogs that are the result of D/L-amino acid changes at nine positions in an 11-amino acid polypeptide, $2^9 = 512$, is given in the literature.[1] A diagram of the 11-amino acid polypeptide is given in Figure 1.

The general analysis problems presented by this data set are how to quickly determine general trends and how to isolate any peculiar results. These general problems lead to the following particular questions: Which are the positions where D/L changes affect potency? Are there combinations of D/L changes that are unusual? and Which polypeptides are not responding as the general trends would predict? We will use two methods that are designed for the examination of large, complex data sets. The first method goes by the name FIRM for formal inference-based recursive modeling.[2,3] At each stage this method does two things. First, categories that are not significantly different from one another are combined. (In the case of the substance P data, each variable, position, has only two categories, so this step is unnecessary.) Second, FIRM uses statistical hypothesis testing to identify the single most important variable for dividing the data set into homogeneous parts. The procedure is recursive and stops when each subgrouping of the data can no longer be subdivided. The internal computations are complex, but the process is easily followed by examination of the analysis of the substance P data set. The second method is visualization, and we use TempleMVV[4] to display all of the data in a way that makes general effects obvious and allows the detailed examination of individual observations in an informative context.

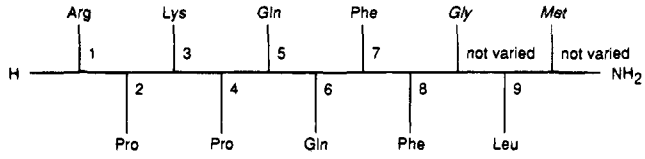

Figure 1. Amino acid sequence of substance P polypeptide. Residues in italics were not varied. Position number is given along the backbone.

In this paper we work systematically through an analytical and a visual examination of the substance P data. There are three points that we hope to make. First, a systematic statistical experimental design is amenable to simple analysis; this well-known characteristic of statistical experimental designs is very important for large data sets. Second, analytical methods can be used not only to determine general features of the data but also to guide a visual analysis. Finally, visual methods can be used to examine detailed features of the data, e.g., detection of unusual data values and effects due to combinations of experimental factors. The analysis process typically involves interplay between analytical and visual methods, but we will present a numerical analysis first followed by a visual analysis. We also examine the question of whether a smaller design could be used to understand the effect of D/L changes.

Statistical Methods

In this experiment there are two possible forms, D and L, of an amino acid at each of nine positions, so there are a total of $2^9 = 512$ distinct chemical entities. This experiment is in the form of a classical statistical experimental design called a full factorial design. The fact that all 512 possibilities are included in the design means that it is a full factorial design. Full factorial designs are well studied and have many useful

* To whom all correspondence should be addressed. Tel: (919) 248-7254. Fax: (919) 248-7645. e-mail: ssy0487@glaxo.com.
† Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108.
® Abstract published in *Advance ACS Abstracts*, June 15, 1995.

0022-2623/95/1838-2784$09.00/0

properties.[5] We remind the reader of two advantages of full factorial designs. The first advantage is called "hidden replication". One way to detect small effects is to use a large number of replications, since with a large number of replications the standard error of a mean decreases with the square root of the number of replicates. In a 2^n factorial experiment, one-half of the observations are at one level of a variable and the other half are at the other. In this experiment, consider position 1, 256 polypeptides have a D-amino acid at this position and the other 256 polypeptides have an L-amino acid at this position. So there is the potential to detect very small effects of changing from D to L at this position. This same argument applies to the other eight positions. It is as if 512 replications were committed to each of the nine questions. If individual experiments had been run to examine each of the positions, then $9 \times 512 = 4608$ observations would be necessary. The full factorial gets the same statistical power with only 512 observations, hence "hidden replication".

It is straightforward to estimate the average effect of a part of a molecule on its biological activity.[6] Unfortunately, the effect of a part of a molecule usually depends upon the other parts that make up the molecule. Parts of molecules interact. The second advantage of factorial experiments is that they can detect the "interaction" of factors. If the effect of changing from D to L changes depending upon what form of amino acid is at another position, the two positions are said to interact. In extreme cases of interaction, the optimum form of an amino acid at one position will change depending upon what is at one or more other positions. If there is interaction and each position is studied and optimized separately, then there is a risk that a suboptimal solution will be found. High-order interactions are interactions of many factors. Full factorial designs can be used to detect high-order interactions.

The detection of high-order interactions is not easy, but equally difficult is explaining them clearly to individuals not trained in the nuances of statistics. Specialized statistical techniques have been developed to automatically find interactions[3,7,8] and graphically display the findings. FIRM[2] is a program for the automatic detection of interactions, and it presents the results of its findings in a tree diagram that contains a wealth of information in a form that is easy to comprehend. The method is recursive and operates in the following way. A set of observations at a node is examined, the variable that can best divide the observations into two or more homogeneous sets is found, and the data in the node are divided into daughter nodes. This process is most easily followed by examination of the results of an analysis. The analysis of the substance P data will be given later.

With large, complex data sets, it is very useful to graph the data in various ways. The point is to take advantage of our visual ability to discover patterns in the data. The visual search is for general effects, specific combinations of variables that act synergistically, and to find unusual observations, statistical outliers, that merit additional scrutiny. We use a software package, TempleMVV, to visually examine this data set.

Experimental Data. Natural amino acids, except glycine, can be in mirror image forms, named dextro (D) and levo (L). The L-form is the natural form. The D/L-configuration and potency of the 512 polypeptides are given in Table 1 of Wang et al.[1] All peptides were tested at a single concentration of 500 nM, and percent inhibition was determined; assay results were between 0 and 100, with larger values better. Careful examination of the original data table revealed that there were two pairs of duplicate polypeptides, ID numbers 102/236 and 229/371. Correspondence with Dr. Wang resulted in the correction of two typographical errors:

	no.	peptide	potency
original	236	RPkPqQfFGLM	6
corrected	236	RPkPqQfFGlM	6
original	229	RPkpQQffGlM	16
corrected	229	RPkpQQffGLM	16

The 11 amino acids of a peptide are given using the single-letter amino acid codes; L-isomers are given in capital letters, and D-isomers are given in lower case letters. We display a few records in Table 1 to emphasize the inherent difficulty of determining general conclusions and unusual data values without the aid of statistical analysis and visualization.

Table 1. Example Peptides and Potencies from Table 1 of Reference 1[a]

no.	peptide	%
1	RPKPQQFFGLM	99
23	RpKPQQfFGLM	91
45	RPKPQQfFGlM	0
67	rPKPqQFfGLM	73
89	RpKPqQFFGlM	8
111	RPKpqqFFGLM	74
133	rpkPQqFFGLM	68
165	rPkPQQfFGlM	0
197	RpkPQqFfGLM	19
229	RPkpQQffGlM	16
261	rpkpQQFFGlM	59
293	rPkpqQfFGLM	42
325	rPKPqQffGlM	2
357	RpKPqqffGLM	9
389	rpkpQqFFGlM	27
421	rPkpqQffGLM	13
453	RpkPQqffGlM	7
485	rpKpqQffGlM	0

[a] Lower case letters indicate the non-natural D-isomers. Note that the 9th and 11th amino acids were fixed in the design.

Supporting information for this paper is available. The 512-observation experiment can be divided into four 1/4 replicates (reps); each rep contains enough information to determine the main features that are related to activity; the methods to construct the 1/4 reps are given. Unusual observations can lead to a misleading analysis; a FIRM analysis is given for each of the 1/4 reps. For one of the reps, the analysis is repeated with outliers removed.

Numerical Analysis Using FIRM

The initial FIRM analysis is given in Figure 2a. First examine the legend. There is a box, and in the box there are statistics describing the observations in the node. These summary statistics include the number of observations, the mean, the standard deviation of the individual observations, and the standard error of the mean. Each box is numbered with a node number. Just above each box the levels of the split variable that are combined to determine the node are given. And on the line that gives rise to the split, the name of the split variable is given along with the p-value for the statistical significance of the split. FIRM reports the p-value as a percent, so a p-value of 0.05 will be reported as 5.0%. With this description of a node, we can now work through the FIRM analysis.

Start at node 1. There are 512 observations, and the mean value for these observations is 27.623. The standard deviation is 28.425. Keep in mind that the observations are percents and can range from 0% to 100%. With the standard deviation about equal to the mean, this data set has a large amount of variation. Some of the variation is variability in the assay, and some is variability induced by differences in the D- and L-forms of the amino acids. The FIRM analysis will attempt to take apart the variability in node 1 and assign it to variations in the positions. As we go down the tree, we should expect to see large differences in the node means and the standard deviations within a node decrease. At the terminal nodes we expect the standard deviation within a node to be similar to the assay variability. Wang et al.[1] state that the variation between duplicate assays is about 5%.

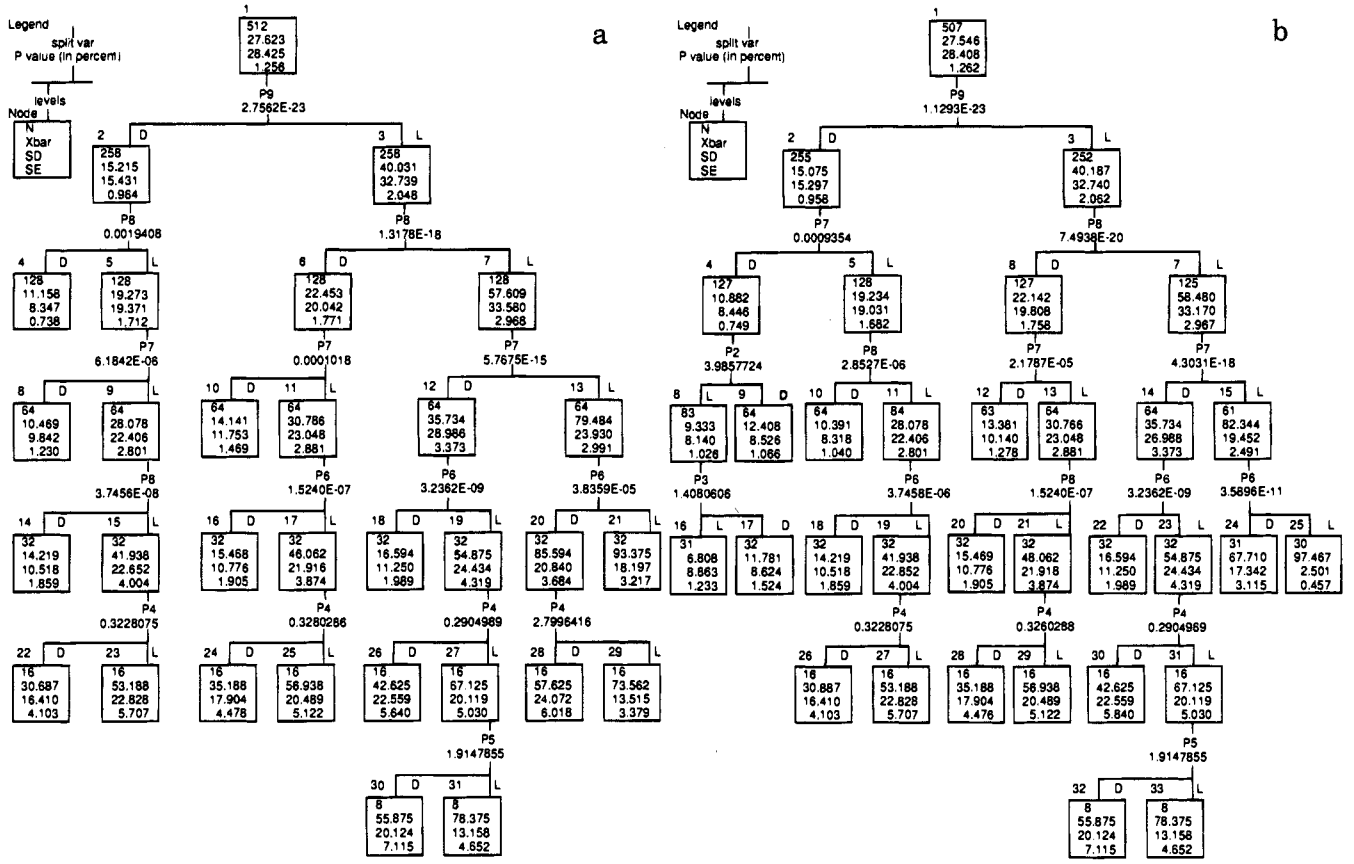

Figure 2. (a) FIRM analysis tree diagram of full factorial. (b) FIRM analysis tree diagram of full factorial, outliers removed.

FIRM finds that position 9 is the most influential position on potency. The p-value of $(2.75 \times 10^{-23})\%$ is highly significant so we can be sure that D/L at position 9 is a very important determinant of potency. The mean of node 2, P9 = D, is 15.215, and the mean of node 3, P9 = L, is 40.031. The standard deviation is decreased in node 2 but has not increased in node 3.

Let us follow the high-potency part of the tree next. Node 3 is split into nodes 6 and 7 based upon position 8. Again, the D-form is less potent, mean = 22.453, and the L-form is more potent, mean = 57.609. The standard deviation is smaller in node 6 but has not decreased in node 7. Notice that position 8 was also used to split node 2. The FIRM algorithm examined all positions for the split of nodes 2 and 3 and did not have to choose the same variable. Notice also that the differences in the means for nodes 4 and 5 and nodes 6 and 7 are quite different from one another:

$$\text{node 5} - \text{node 4} = 19.273 - 11.156 = 8.117$$

$$\text{node 7} - \text{node 6} = 57.609 - 22.453 = 35.156$$

So the size of the effect induced by changing from D to L at position 8 changes as a function of what is at position 9. From Figure 1 we see that these positions are beside each other, so it is not surprising that an interaction might occur.

We see that FIRM did not split node 4. Nodes 5–7 were each split using position 7. At this level we see a rather consistent decrease in the standard deviation at each node relative to the standard deviation at node 1. You can trace down each branch of the tree examining means, standard deviations, and differences between the means for pairs of nodes.

Note that at each split the sample size in the descendant nodes decreases. Because we are analyzing a factorial design, equal numbers of observations split into the descendant nodes. Because the number of observations gets smaller as FIRM proceeds with its analysis, the statistical power is expected to decrease. The variability of a node is expected to decrease as we go down the dendrogram, and this decrease in variability will increase statistical power. But the decrease in sample size is typically more rapid than the decrease in variability so, in general, the statistical power decreases as we go down the dendrogram. Ultimately FIRM will be unable to split nodes, and the process stops.

If interest is in high-potency polypeptides, then Figure 2a can be used to give a simple decision rule: Positions 9–6 should be set to L; positions 4 and 5 can be important; for positions 1–3, the differences between D and L are relatively unimportant.

Two additional comments. First, the potencies were determined at a single dose, are percents, and are limited to be between 0% and 100%. So once the potency achieves 100%, the full beneficial effect due to changes at the other positions may not be apparent. Second, it is of interest to see how the variability of the observations changes across the terminal nodes; there may be outliers in the data set that have not been identified in the splitting process. We will come back to that question.

Graphical Analysis Using TempleMVV

There are literally millions of ways that high-dimension data may be graphed; an important question is which views are likely to be informative. The FIRM

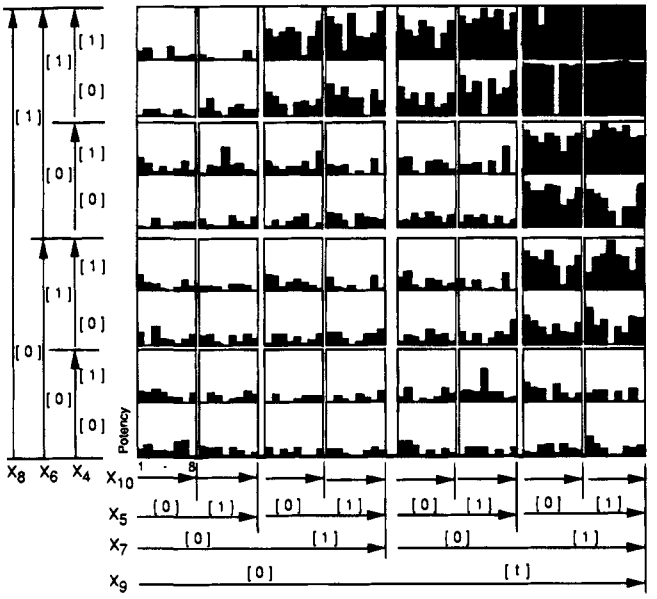

Figure 3. Multidimensional diagram. Within each cell, potency versus X_{10}, an index of positions 1–3. Rows and columns index positions 9–4.

analysis indicates that positions 9–4 are important, so we should graph results with respect to those variables. Positions 1–3 do not appear to be important, so variation induced by those variables might be considered random. The graphical display of the data should allow examination of the effect of positions 9–4 and an assessment of the randomness of positions 1–3. To that end, we created nine variables, X_1–X_9, that take the values 0/1 depending upon whether the amino acid at positions 1–9 is D and L. Next we created a new variable, X_{10}, that takes the values 1–8 depending upon the values of X_1–X_3 as given in the following table:

X_{10}	X_1	X_2	X_3
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

X_1–X_{10} will be used in a complex graph which we will now explain. Examine Figure 3. Each small box shows a vertical bar graph of potency versus X_{10}. As X_{10} ranges from 1 to 8, the potency is thought to be random, so we expect to see no discernible pattern within each box. The rows and columns of boxes are indexed by variables X_9–X_4 is a nested fashion. FIRM indicated that position 9, X_9, was most important, and X_9 is displayed along the bottom of Figure 3. The non-natural D-amino acids are in columns 1–4, and the natural L-amino acids are in columns 4–8. The bars are generally taller in columns 5–8, visually confirming the FIRM analysis and indicating higher potency for the natural amino acid in position 9. Rows 1–4 are associated with $X_8 = 1$, the natural amino acid in position 8. Rows 5–8 correspond to the non-natural amino acid at position 8. Again, there is visual confirmation of the FIRM analysis; the most potent polypeptides are predominantly in the first four rows. Variables X_7 and X_5 are nested within X_9 and index the columns. Variables X_6 and X_4 are nested within X_6 and index the rows. The general effects of variables X_4–X_9 can be examined by comparing the heights of the bars in the rows and columns.

Figure 3 also facilitates the examination of fine detail of the data set. For example, look at the next to the last small box in row 2. All of the bars are high except for the 5th one, which is essentially 0. Something appears to be amiss. Either the compound was misynthesized, the assay is incorrect for some reason, or there is some very unusual effect going on. In any case, this result merits examination by the experimenters. There are other unusually low potencies; see row 1 columns 3–5 and 7; see row 2 column 6. There appears to be an unusually large potency in row 7 column 6. The power of this display is that general effects can be observed, and with the figure arranged by the general effects, it is quick and easy to find unusual observations in the data set.

FIRM Analysis with Outliers Removed

Figure 3 indicates the presence of several observations that appear very inconsistent with the rest of the data. Five observations

ID1	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_6	X_9	Y	result
56	0	1	0	1	0	1	1	1	1	64	too low
183	0	1	1	1	1	0	0	0	1	62	too high
184	0	1	1	1	1	0	0	1	0	51	too high
187	1	0	0	0	0	1	1	1	1	0	too low
188	1	0	0	0	1	0	1	1	1	0	too low

were removed, and the FIRM analysis was rerun (see Figure 2b). There were several changes relative to Figure 2a. The mean in node 25, Figure 2b, increased to 97.5 from 93.4 in node 21, Figure 2a. The standard deviation also decreased dramatically, from 18.2 to 2.5. In Figure 2a, node 2 was split by position 8. With the outliers removed, this node was split by position 7. Also in Figure 2b, node 4 was split by position 2 and node 8 by position 3, and for each of these two splits the non-natural amino acid was higher in potency. The *p*-values for these splits were 4.0% and 1.4%, respectively, so the splits might be type I errors, the result of chance.

Is a Smaller Design Possible?

It is interesting to consider if a smaller experiment could have determined the general trends discovered in this large experiment. With results for all $2^9 = 512$ observations, it is possible to select out four "$^1/_4$ rep" fractional factorial designs, each having $2^{(9-2)} = 128$ observations. (The method of construction of a $^1/_4$ rep design is given in the supporting information.) In each $^1/_4$ rep, it is possible to estimate all main effects and two-way interactions.

Each of four $^1/_4$ reps was selected, and a FIRM analysis was run. (The FIRM dendrograms are available in the supporting information.) For each $^1/_4$ rep, position 9 was found to be most important. Positions 7 and 8 form the next level splits. The remaining splits usually involve positions 6 and 4. Each dendogram is somewhat different in the order of positions that are identified, but in general the same positions are determined to be important as were found in the analysis of the complete data set. So a much smaller experiment

could have been used to detect the general trends induced by changing L- to D-isomers. It is interesting to note that high-order interactions are used to divide the 512 observations into the four $^1/_4$ reps, and for this to be sensible, it is assumed that these interactions are negligible. It appears from the full analysis that high-order interactions are present, so the fractional factorial, though informative, could give erroneous results. In this case, positions 9–6 were all identified as important in each FIRM analysis even though the relative importance of the positions changed with each $^1/_4$ rep.

Discussion

FIRM and TempleMVV clearly pick out the important positions where D/L-substitutions affect the potency of the substance P polypeptide. Positions 9–6 are very important; positions 5 and 4 are only important when potency is low (any of positions 9–6 are D). The FIRM display is particularly effective in allowing quantitative comparisons over a complex data set. Once the graphical display is set up on the basis of the FIRM analysis, it displays the general effects and allows the examination of fine details of the data set. This examination of fine detail is very important as it greatly facilitates finding unusual data values, outliers. These outliers are important for two reasons. Incorrect values can corrupt analysis, so they need to be identified and removed. It is often essential to recompute the analysis after outliers are removed. On the other hand, the outliers may not be incorrect; correct but unusual data values often are the starting point for discovery. In the case of the substance P data set, one of the goals of the research was to identify peptides that were potent but had a high proportion of D-amino acids. D-Amino acids are less likely to be degraded by the body and hence are more likely to be better starting points for drug development. It is clear from a cursory examination of the raw data that high-potency polypeptides can be made by using natural L-amino acids. Both the numerical analysis and the examination of outliers in this data set point to polypeptides that are possible candidates for high-potency/high D-containing peptides. The peptide with L in positions 7–9 and D in positions 1–6 had a potency of 86%. The peptide with L in positions 6–9 and D in positions 1–5 had a potency of 96%. These peptides are potential starting points for further development.

As complex as this data set is, see Table 1, it is still very simple in many respects. Each variable takes only two values; positions are either L or D. All possible combinations of L/D are in the data set, so it is possible to examine all possible interactions. If there were more chemical components at each position and the components were described with many numerical descriptors, then the analysis problem would be more difficult (and realistic). The problem would be much more difficult (impossible?) if the set of compounds was some sort of catch-as-catch-can collection. It remains to be seen how this general approach, find the important variables with FIRM and then use the results of that analysis to direct interesting views of the data with TempleMVV, would work in more complex situations.

Once the layout of the TempleMVV display is understood, finding unusual values appears and is very simple. This simplicity is deceptive. Detecting outliers in complex multivariate data sets can be very difficult. The presence of multiple outliers can greatly confuse standard analysis. Because of the data-splitting nature of FIRM, outliers are only in their own branch of the tree. For that reason, a FIRM analysis can be somewhat less confused by outliers. Outlier analysis typically proceeds in two steps. First, the general features of the data are modeled and subtracted out. The resulting "residuals" are then compared to some measure of variability. Multiple outliers can cause the general features of the data to be poorly modeled. Also, the unbiased measurement of variability is difficult with single or multiple outliers. Analysis in the presence of outliers is typically rather *ad hoc* and usually iterative. The data are modeled and residuals computed. The presumed outliers are removed, and the analysis is repeated.

The binding of these polypeptides is very nonadditive. For example, the effect of the D/L change at position 8 is highly dependent on position 9. The analysis method should be effective in the presence of interactions. The FIRM methodology was designed to detect interactions. Often the binding of compounds to receptors is nonadditive, so FIRM is a potentially useful analysis method in other complex situations.

FIRM and TempleMVV work together synergistically; the FIRM results are used both to organize the TempleMVV display and to identify the apparently irrelevant predictors defining the "noise" variable X_{10}, and then we use the TempleMVV display to identify outliers and refine the FIRM analysis by rerunning the analysis without outliers in the data set. The general and particular features of this complex data set are quickly and easily identified.

Supporting Information Available: Method for construction of a $^1/_4$ rep design and FIRM dendograms (6 pages). Ordering information can be found on any current masthead page.

References

(1) Wang, J.-x.; DiPasquale, A. J.; Bray, A. M.; Maeji, N. J.; Geysen, H. M. Study of Stereo-Requirements of Substance P Binding to NK1 Receptors using Analogs with Systematic D-Amino Acid Replacements. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 451–456.
(2) Hawkins, D. M. *FIRM Formal Inference-based Recursive Modeling, Release 2*; University of Minnesota: St. Paul, MN, 1994.
(3) Hawkins, D. M.; Kass, G. V. Automatic Interaction Detection. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press: Cambridge, U.K., 1982; pp 269–302.
(4) *TempleMVV Version 1.3*; Temple University and Mihalisin Associates, Inc.: Ambler, PA, 1994.
(5) Box, G. E.; Hunter, W. G.; Hunter, S. *Statistics for Experimenters*; John Wiley & Sons: New York, 1978.
(6) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
(7) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
(8) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.

JM950036F