A COMPUTER SYSTEM FOR THE ANALYSIS OF DATA GENERATED BY MOLECULAR STUDIES OF DNA

by

Scott McCourt
Computer Science Department
Indiana University
Bloomington, Indiana 47405

TECHNICAL REPORT NO. 85

A COMPUTER SYSTEM FOR THE ANALYSIS OF DATA GENERATED BY MOLECULAR STUDIES OF DNA

SCOTT MCCOURT

MAY, 1979

Submitted to the faculty of the Graduate School in partial fulfillment of the requirements for the degree Master of Science in the Department of Computer Science, Indiana University.

*Research reported herein was supported in full by the National Institute of Health under a grant numbered AI-13382.

ACKNOWLEDGEMENTS

I would like to thank Dr. Stefan Surzycki of the Department of Biology for making the project possible and for his many helpful discussions during its development. I also thank the graduate students in his lab for providing the actual data without which, the testing of the programs would have been impossible. I would also like to thank my thesis advisor, Dr. Cynthia Brown, for her patient reading of several drafts of this thesis and my wife Maureen for her patience throughout the project.

TABLE OF CONTENTS

INTRODUCTION:

The DNA Analysis System described here is an implementation of a method developed in our lab to analyze data generated by electron microscopic studies of DNA. The purpose of this paper is to describe, in general, the method we use to accomplish this analysis and also to describe, in particular, the capabilities of our specific implementation. While some of the details of these two programs are specifically tailored to our hardware configuration and will therefore be of little interest to those with other configurations, the general method which we employ will surely be of interest to those with similar data to analyze.

Our system consists of two interactive computer programs, DNA1 and DNA2. DNA1 is used to maintain disk files containing experimental data and allows storage, retrieval and modification of the data. DNA2 is used to analyze the stored data according to user parameters.

These programs are written in the BASIC programming language and are implemented on an Apple II microcomputer. The system contains 48K of user memory and runs under the Apple supplied Disk II Floppy Disk Operating System (DOS). Printed output is produced on an Integral Data Systems line printer with tractor feed and graphics options.

DATA DESCRIPTION:

The data which we analyze on our system is generated as follows. An experiment is run whereby certain protein molecules are caused to bind to specific regions of DNA molecules. The general function of these specific sites is well known but their exact location on the DNA molecule has not been determined. Once these proteins have been attached to the DNA, the DNA is prepared for viewing on an electron microscope. In a photograph, the DNA-protein complex resembles a thread (DNA) with small spheres attached (proteins). These small spheres are commonly called bushes because of their appearance when enlarged. The DNA molecules are measured using an electronic digitizer (Numonics Corporation) and the relative positions of the bound proteins are recorded as well as the total length of the DNA molecule. The molecules are measured twice to prevent large measuring errors. A sample number is assigned to each molecule as it is measured and it is stored in a disk file using DNA1, along with the total length of the DNA molecule and the relative positions of the bound proteins which are stored as percentages of the total length. Photographs are enlarged to different degrees and therefore absolute measurements have no real meaning.

The function of DNA2 is to determine, if possible, where the specific sites of attachment are located. It will be shown in the next section why this is not as straightforward a process as it first appears to be.

GENERAL OVERVIEW:

DNA1 allows the user to enter new data into a disk file or to retrieve data which has previously been save. Once stored in a disk file, the user has the option of 1) adding new samples to the file, 2) deleting samples from the file, 3) printing out the data file on the line printer, and 4) storing the data back to the disk.

A data description is stored along with the experimental data to identify the data set once it has been read from the disk. Users are also encouraged to give their files descriptive filenames to facilitate easy access to specific data sets.

Error handlers have been included to prevent errors in entering the experimental data and to handle disk I/O errors (disk full, incorrect filename and redundant filenames). These features are illustrated in a later section of this paper.

The function of DNA2 is to analyze the data stored by DNA1 in an attempt to determine, if possible, where the specific sites of attachment of the protein molecules to the DNA are located. DNA2 allows many options when analyzing the experimental data to aid the user. The method we use is as follows.

Each DNA molecule may be divided into n equal lengths, where n is decided by the user. A histogram can then be generated from this data with intervals 1-n on the x-axis (typically 1-100% in 1% increments) and the number of molecules with a bush present in a given interval on the y-axis. This relatively straightforward process is complicated by the fact that for a given molecule viewed in a photograph, it is not known which end is the left end and which is the right end. Therefore any one molecule has two possible orientations. Furthermore, each molecule may have any number of proteins bound and a small amount of random binding occurs, contributing to the "noise" which appears in a histogram. DNA2 takes as its input data with no apparent ordering, and attempts to generate the "best" histogram it can by manipulating the orientation of molecules in the data set. We define the "best" histogram as one with the tallest peaks and deepest valleys (least amount of "noise"). We use the following coefficient to evaluate different arrangements of the same data set:

$$C_H = \sum_{i=1}^{n} H(i)^2 \; / \; m$$

where n is the number of equal intervals in the histogram, H(i) is the number of molecules with a bush in interval i, and m is the total number of bushes represented in the histogram. The inclusion of m allows us to compare arrangements of different data sets.

If we had only a small number of molecules in the data set, we could easily generate the maximum C_H by iteratively trying all the possible arrangements of the samples within the data set. While this would be useful for very small data sets, it is not practical for larger data sets which we typically analyze (75-100 samples). We try to approximate that solution using a much faster technique which will now be described.

The first step is to compare molecules on a one-to-one basis in order to determine which molecules are similar to each other (i.e., which molecules have bushes within (or near) the same interval H(i)). Each possible pair of molecules may be compared as follows:

1) For the first molecule, generate an array H, such that for each element H(k), $1 \leq k \leq n$, H(k)= the number of bushes in interval k.

2) For each bush on the second molecule within interval j, calculate the following:

$$C_F = \sum_{i=1}^{n} H(j) + .3413 \times (H(j-1) + H(j+1)) + .1359 \times (H(j-2) + H(j+2))$$

where n is the number of bushes on the molecule and j refers to the interval in which bush i is found. This calculation compares a bush in an interval on one molecule to the corresponding interval on the other molecule as well as to the 2 intervals on either side of it. Matches in these surrounding regions are not counted as heavily as a "direct hit" (a match in the same interval).

We have derived this formula to be consistent with the fact that many occurrences of a bush at aspecific site, when measured, will be found to form a normal distribution around a central site. Since the electron microscope results allow us to measure the position of a bush to within 2% of the length of the DNA, we form a hypothetical normal distribution around the bush we have actually measured. This also forms the basis of our smoothing function used in DNA2 to produce histograms of the results of a given analysis.

3) Reverse the orientation of the second molecule and generate C_R, the reverse correlation.

4) These correlations as well as their difference can be compared to those for other pairs of molecules in order to determine which molecules are the most similar.

Once we have an idea of which molecules in the data set are most similar, we pick a small number of those (typically 5-8) and iteratively manipulate their orientation to yield an arrangement with the highest possible value for C_H. An explanation of how we pick these samples is given in the section "Using DNA2". We refer to this process as the "best histogram fit".

These molecules are entered into an array H with intervals 1-n. The remaining molecules are now added to this histogram in the orientation which yield the higher of the values, C_F and C_R. Options have been included in DNA2 to restrict the conditions under which a sample is added to the developing arrangement based on the value $|C_F - C_R|$. The higher this value, the more likely it is that the current sample belongs in the arrangement in either the forward or reverse orientation.

Once the analysis is finished, the resulting histogram may be viewed on the video monitor and printed on the line printer. Statistics may be performed on the results and the arranged data may be saved in a disk file for later reference.

DATA STRUCTURE:

The data structure for these programs has been kept simple in order to conserve space but at the same time allow for quick access to individual samples. It consists primarily of 4 arrays, A, B%, I and S.

Array I contains the sample numbers of all the samples contained in the file. To locate an individual sample, array I is scanned until the corresponding sample number is found. The index, X, which corresponds to this sample is then used to index into array B%. The value obtained from array B% points to a value in array A, which is the first item in the sample. Therefore to access all the items for a given sample whose identifying number is found in I(X), one need only access those items in array A indexed from B%(X) to B%(X+1) - 1. The fourth array, S, consists of the actual lengths of all the samples.

USING DNA1:

DNA1 allows the user to create and modify data files which will be analyzed later by DNA2. Instructions on how to use DNA1 are divided up into the following categories:

1) Creating new files

2) Manipulating old files

   a) Adding data

   b) Deleteing data

   c) Printing a file

   d) Saving a data file

3) Error checking provided by DNA1

1) CREATING A NEW FILE

To create a new data file, the user responds with 'NEW' to the question which asks whether he will be working with new or old data. After a short list of instructions, he is asked to input a short data description (up to 40 characters) which will be stored along with the sample data. A longer description causes an error message to be printed. After entering the data description, the user will see the message:

```
BEGIN  DATA ENTRY:

SAMPLE #?
```

A user-assigned sample number is then input and the program asks for the first measurement for that sample. Measurements are input as pairs (obtained by measuring the samples twice) and are averaged by the program to yield a single value. The user continues to input measurements, corresponding to the positions of the bound proteins, in ascending order and adds the actual length measurements last. After the length measurements have been entered, the user enters "0,0" to signify the end of a sample. The binding sites are then converted to percentages of the total length. The program will now ask for the next sample number. Samples are entered in this manner until finished. When data entry is finished, entering a zero (0) for the sample number will terminate data entry.

At this point we advise the user to save his data on the disk (see below) to prevent accidental data loss. This data is now considered to be old data and may be manipulated as explained below.

2) MANIPULATING AN OLD FILE

A user accesses old data (already on the disk) by typing 'OLD' in response to the question as to whether he will be working with NEW or OLD data. After reading old data from a disk file, the user has four options:

a) add sample(s)

b) delete sample(s)

c) print all samples

d) save data on the disk and stop

A) ADD SAMPLE(S) (type "A" when given options)

Samples are entered in the same manner as when creating a new file (see above). Type a zero (0) for the sample number to terminate data entry. After terminating data entry the user has the option of ordering the entire file so that the samples are put into descending order in terms of the number of bushes present. The usefulness of this option is explained in the description of how to use DNA2.

B) DELETE SAMPLE(S) (type "D" when given options)

Sample numbers are entered one at a time to delete samples. If a sample is not found in the file, a message will be printed to indicate that fact. Enter a zero to terminate sample deletion.

C) PRINT SAMPLES (type "P" when given options)

This option causes a list of all the samples to be printed either on the video monitor or on the line printer (hardcopy). For each sample the sample number, sample length and position of bound proteins as a percentage of total length are printed.

D) SAVE DATA (type "S" when given options)

The data file will be stored on the disk under the filename provided by the user. If the filename given is already present on the disk, a message is printed to prevent the user from accidentally overwriting another file.

3) ERROR CHECKING PROVIDED BY DNA1

The program has been designed to anticipate certain types of errors and prevent them. This error checking takes place at data entry.

When adding data, three types of error checking are performed. First, the sample number of a new sample is checked against those already appearing in the file to prevent the presence of redundant samples.

Second, the measurements for the bushes must be entered into the program in ascending order. This motivates the user to keep his data in an organized fashion and attempts to prevent errors when entering bush measurements. The third type of error checking has been found to be very useful. As pairs of measurements are entered for a sample, their difference is calculated. If the difference is greater than five, a message is printed indicating that fact and the user is asked:

```
IS THAT O.K. (YES,NO)?
```

This prevents the common error of displaced decimal points. Our pairs of measurements tend to be very close so we have arbitrarily chosen 5 as our cutoff for printing a message. For other types of data, this value can be easily changed.

What follows are some annotated sample outputs which demonstrate all of the previously stated conventions and behaviors of DNA1. These, in combination with the previous description should enable any user to easily use the program for his own data.

```
SAMPLE OUTPUTS FROM DNA1:

RUN
-----------------------------------------
                DNA1
-----------------------------------------

PROGRAM FOR DATA ENTRY, RETRIEVAL AND
MODIFICATION

NOTE: EACH FILE IS LIMITED TO 200
      SAMPLES

-----------------------------------------

NEW OR OLD DATA (NEW,OLD)?new              We'll start by creating a new
                                           data file
-----------------------------------------
       ***** INSTRUCTIONS *****
-----------------------------------------
TO TERMINATE DATA ENTRY, ENTER 0 FOR       Short instruction summary
SAMPLE #

TO START NEXT SAMPLE, ENTER 0,0

TO DELETE CURRENT SAMPLE, ENTER 999,999

-----------------------------------------

INPUT DATA DESCRIPTION (ONE LINE ONLY)

?test run for illustration of dna1 program

** TOO LONG **                             Data description too long

INPUT DATA DESCRIPTION (ONE LINE ONLY)

?test run for illustration of dna1         Shorter data description

-----------------------------------------
BEGIN DATA ENTRY:
-----------------------------------------

SAMPLE #?1000                              Unique sample number

MEASUREMENT 1?1.25,1.28                    Measurements are entered as
MEASUREMENT 2?34.19,35.26                  pairs
MEASUREMENT 3?87.8,89.8
MEASUREMENT 4?0,0                          End of sample
```

```
------------------------------------

SAMPLE #?1001

MEASUREMENT 1?1.28,13.8                     Error in entering data
                                            (displaced decimal point)
** WARNING:  THE DIFFERENCE BETWEEN THE
LAST TWO ENTRIES IS GREATER THAN 5

IS THAT O.K. (Y,N)?n                        We can change it by responding
                                            no
MEASUREMENT 1?1.28,1.38                     Corrected entry
MEASUREMENT 2?45.65,47.67
MEASUREMENT 3?88.2,87.3
MEASUREMENT 4?0,0

------------------------------------

SAMPLE #?1002

MEASUREMENT 1?34.2,35.7                     We left out one bush measurement
MEASUREMENT 2?999,999                       999,999 deletes this sample

** SAMPLE DELETED **

------------------------------------

SAMPLE #?1002

MEASUREMENT 1?5.65,5.78
MEASUREMENT 2?34.2,35.7
MEASUREMENT 3?87.9,88.34
MEASUREMENT 4?0,0

------------------------------------

SAMPLE #?0                                  Terminate data entry

------------------------------------

DO YOU WANT THE DATA PUT IN ORDER           We'll put the data in order
(Y,N)?y                                     this time

------------------------------------

TO MODIFY THIS FILE YOU MAY:                Option list

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE
     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP

------------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?p
```

Before we save the data we'll print it on the screen to check it

```
------------------------------------
HARDCOPY (Y,N)?n

------------------------------------
FILENAME:

------------------------------------

TOTAL NUMBER OF MOLECULES = 3

MOLECULES WITH 2 BUSHES = 3

------------------------------------
SAMPLE # 1000   SIZE= 88.8
1.4  39.1
------------------------------------
SAMPLE # 1001   SIZE= 87.75
1.5  53.2
```

Sample #, size and bush positions are given for all samples

```
------------------------------------
SAMPLE # 1002   SIZE= 88.12
6.5  39.7
------------------------------------

TO MODIFY THIS FILE YOU MAY:

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE

     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP

------------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?s
```

The data looks O.K. so we'll store it on the disk under the filename 'TESTDATA'

```
------------------------------------
WOULD YOU LIKE TO SAVE THIS DATA
(Y,N)Y

FILENAME?testdata
```

```
DATA HAS BEEN SAVED

FILENAME= TESTDATA
```

```
RUN
----------------------------------------
                DNA1
----------------------------------------

PROGRAM FOR DATA ENTRY, RETRIEVAL AND
MODIFICATION

NOTE: EACH FILE IS LIMITED TO 200
      SAMPLES

----------------------------------------

NEW OR OLD DATA (NEW,OLD)?old                    Our file, Testdata, is now
                                                 considered old data
FILENAME?tesdata                                 Filename spelled incorrectly

TESDATA NOT FOUND ON DISK

CATALOG LISTING (Y,N)?n                          We could get a catalog listing
                                                 if we couldn't remember it

FILENAME?testdata
----------------------------------------
DATA DESCRIPTION:

TEST RUN FOR ILLUSTRATION OF DNA1                This is the right data

----------------------------------------

TO MODIFY THIS FILE YOU MAY:

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE

     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP

----------------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?d                                      We want to delete sample #1002

----------------------------------------
ENTER SAMPLES TO BE DELETED BELOW:


SAMPLE #?10002                                   We typed it incorrectly
SAMPLE NOT FOUND

SAMPLE #?1002                          Now it's right

SAMPLE #?0                             Terminate sample deletion

-----------------------------------

TO MODIFY THIS FILE YOU MAY:

    1) ADD SAMPLES TO THE FILE

    2) DELETE SAMPLES FROM THE FILE

    3) PRINT OUT THE DATA

    4) STORE THE DATA ON THE DISK
       AND STOP

-----------------------------------

ADD, DELETE, PRINT OR STOP
(A,D,P,S)?a                            Add another sample

-----------------------------------

BEGIN DATA ENTRY:
-----------------------------------

SAMPLE #?1005

MEASUREMENT 1?2.34,2.45
MEASUREMENT 2?46.7,47.89
MEASUREMENT 3?65.78,67.98
MEASUREMENT 4?87.9,88.7
MEASUREMENT 5?0,0

-----------------------------------

SAMPLE #?0                             Terminate data entry

-----------------------------------

DO YOU WANT THE DATA PUT IN ORDER
(Y,N)?y                                Put the data in order before
-----------------------------------    saving it
TO MODIFY THIS FILE YOU MAY:

    1) ADD SAMPLES TO THE FILE

    2) DELETE SAMPLES FROM THE FILE

    3) PRINT OUT THE DATA
```

```
     4) STORE THE DATA ON THE DISK
        AND STOP

----------------------------------

ADD, DELETE, PRINT OR STOP
(A,D,P,S)?p                              Print the data out first
----------------------------------

HARDCOPY (Y,N)?n

----------------------------------

FILENAME: TESTDATA

----------------------------------

TOTAL NUMBER OF MOLECULES = 3

MOLECULES WITH 2 BUSHES = 2
MOLECULES WITH 3 BUSHES = 1

----------------------------------
SAMPLE # 1005   SIZE= 88.3
2.7  53.6  75.7
----------------------------------
SAMPLE # 1000   SIZE= 88.8
1.4  39.1
----------------------------------
SAMPLE # 1001   SIZE= 87.75
1.5  53.2
----------------------------------

TO MODIFY THIS FILE YOU MAY:

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE

     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP

----------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?s                              Save the data
----------------------------------


WOULD YOU LIKE TO SAVE THIS DATA
(Y,N)?y
```

```
FILENAME?testdata

TESTDATA ALREADY IN CATALOG

OVERWRITE (Y,N)?y

DATA HAS BEEN SAVED

FILENAME= TESTDATA
```

This prevents us from accidentally writing over our or someone else's file which has the same name

```
RUN
------------------------------------
               DNA1
------------------------------------

PROGRAM FOR DATA ENTRY, RETRIEVAL AND
MODIFICATION

NOTE: EACH FILE IS LIMITED TO 200
      SAMPLES

------------------------------------

NEW OR OLD DATA (NEW,OLD)?old

FILENAME?testdata
------------------------------------
DATA DESCRIPTION:

TEST RUN FOR ILLUSTRATION OF DNA1

------------------------------------

TO MODIFY THIS FILE YOU MAY:

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE

     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP

------------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?a                              We want to add samples #1005
                                         and #1006
------------------------------------
BEGIN DATA ENTRY:
------------------------------------

SAMPLE #?10005                           Wrong sample number
                                         We can delete it
MEASUREMENT 1?999,999

** SAMPLE DELETED **

------------------------------------

SAMPLE #?1005                            Now it's right
SAMPLE #1005 ALREADY IN FILE.

RE-ENTER
```

Sample #1005 has already been entered so we don't have to enter it

---

```
SAMPLE #?1006
MEASUREMENT 1?3.45,3.56
MEASUREMENT 2?45.67,46.75
MEASUREMENT 3?87.7,89.9
MEASUREMENT 4?0,0
```

Enter sample #1006

```
SAMPLE #?0
```

Terminate data entry

```
DO YOU WANT THE DATA PUT IN ORDER
(Y,N)?y
```

```
TO MODIFY THIS FILE YOU MAY:

     1) ADD SAMPLES TO THE FILE

     2) DELETE SAMPLES FROM THE FILE

     3) PRINT OUT THE DATA

     4) STORE THE DATA ON THE DISK
        AND STOP
```

```
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?p
```

Print out the data

```
HARDCOPY (Y,N)?n
```

```
FILENAME: TESTDATA
```

```
TOTAL NUMBER OF MOLECULES = 4

MOLECULES WITH 2 BUSHES = 3
MOLECULES WITH 3 BUSHES = 1
```

```
------------------------------------
SAMPLE # 1005   SIZE= 88.3
2.7  53.6  75.7
------------------------------------
SAMPLE # 1000   SIZE= 88.8
1.4  39.1
------------------------------------
SAMPLE # 1001   SIZE= 87.75
1.5  53.2
------------------------------------
SAMPLE # 1006   SIZE=88.3
4  52.3
------------------------------------

TO MODIFY THIS FILE YOU MAY:

     1 ) ADD SAMPLES TO THE FILE

     2 ) DELETE SAMPLES FROM THE FILE

     3 ) PRINT OUT THE DATA

     4 ) STORE THE DATA ON THE DISK
         AND STOP

------------------------------------
ADD, DELETE, PRINT OR STOP
(A,D,P,S)?s                                     Save the data on the disk
------------------------------------


WOULD YOU LIKE TO SAVE THIS DATA
(Y,N)?y

FILENAME?testdata

TESTDATA ALREADY IN CATALOG

OVERWRITE (Y,N)y

DATA HAS BEEN SAVED

FILENAME= TESTDATA
```

USING DNA2:

DNA2 allows the user 3 major options with a number of minor options within each of them. The major options include:

1) Correlations - to determine which samples in the data set are similar to each other.

2) Arrangement - to arrange the user's experimental data based on his selected parameters.

3) Histograms - to provide a graphic representation of the results of an individual arrangement and statistics based on those results.

These major options and the minor options within each will be described here. For hints on how each may be used see the "Sample Outputs of DNA2" section of this paper.

1) CORRELATIONS:

The purpose of this option is to identify similarities between samples. This is done so that a small number of similar samples may be chosen as a model, against which the remaining samples may be arranged. For a pair of samples, C_F, C_R and $|C_F - C_R|$ are calculated. Since these values are dependent on the number of bushes present on the samples, C_F/n, C_R/n and $|C_F - C_R|/n$ are also calculated where n is the number of bushes on that sample in the pair which has the smaller number of bushes. These values are printed out for all possible pairs of samples or only a subset depending on the parameters selected by the user.

Options Available When Obtaining Correlations:

a) Interval Size - this parameter defines the number of intervals into which the user wishes to have each sample divided. Our data typically uses an interval size of 1% (100 intervals). The larger the interval size, the less strict the program will be in deciding that 2 samples are similar.

b) Minimum and Maximum Bush Number - this parameter allows the user to put a restriction on the samples to be included when obtaining correlations for a given data set. This is used to shorten the time which it takes to run an entire set of correlations. Since we will only need to know which samples are most similar, it is usually advisable to leave out those samples which only have a small number of bushes.

c) Minimum Correlation For Printout - since the correlation values are printed on the line printer, this parameter allows the user to have only those correlations which fall above a certain value printed out. The value which is tested is $\left|C_F - C_R\right|$. After using the program a few times, the user will be able to predict the magnitude of the correlations for his data and reasonably pick a value to limit the amount of output produced but at the same time generate the needed results.

d) Miscellaneous Features - if the user decides to terminate the running of his correlations abnormally, typing an "S" will terminate the procedure and allow him to start again. This is so the user does not have to wait for all the correlations to be calculated in the event that he has made an error in choosing parameter values.

Once the required values have been obtained, the user has the option of printing out samples on the printer to see which samples the program has chosen as being similar. After answering "Y" to the question:

WOULD YOU LIKE CERTAIN MOLECULES PRINTED OUT (Y,N)?

the user need only enter one sample number after another. After inputting all of the sample numbers which he wishes to have printed, entering a zero (0) will cause all of those samples to be printed on the video monitor and on the line printer. This process will aid the user in choosing those samples which will be used as a model, against which to arrange the other samples (see "Arrangement" below).

2) ARRANGEMENT:

The purpose of this option is to accomplish the actual arrangement of the samples in such a way as to give the user results which will allow him to predict the location of the sites he is investigating. This option contains a large number of minor options in order to allow a great amount of flexiblity when arranging the data. Along with this flexibility will come some confusion as to how these options may be used to the user's advantage. The options are explained below and examples are give as to how certain options may be used in combination. Any confusion which exists after reading about these options should be cleared up in the section "Sample Outputs From DNA2".

Options Available When Performing an Arrangement

a) Interval Size - this option serves the same function as it does when obtaining correlations (see above).

b) Putting Certain Molecules First - since the data will be arranged beginning with the first samples in the data array, the user will want to put the similar samples, determined by the correlations, at the top of the data set. After responding "Y" to the question:

```
DO YOU WANT TO PUT CERTAIN MOLECULES
FIRST (Y,N)?
```

the program asks:

```
HOW MANY MOLECULES (1-10)?
```

After responding to this question, the sample numbers of the desired samples are entered one at a time, causing them to be moved in the corresponding order to the top of the data array.

c) Best Histogram Fit - once the user has moved his similar samples to the top of the data set using option (b) above, he may then perform a Best Histogram Fit (seeGeneral Overview) on up to 10 of the first samples in the data set. This arrangement of the first few samples will serve as a model against which to arrange the remaining members of the data set.

d) Minimum and Maximum Bush Number - as when obtaining correlations (1b above), this option allows the user to put a restriction on those samples which will be included in forming the final arrangement. This allows the user to leave out those samples with a small number of bushes. Those samples with a large number of bushes are more likely to yield better results. For this reason, the user is allowed to order his samples in decreasing order in terms of bush number in DNA1 (see "Add Samples" under "Using DNA1"), thereby causing samples with the most bushes to be added to the arrangement first.

e) Correlation Cutoff Value - as each sample is added to the developing data arrangement the values C_F/ (N) (M) and C_R/ (N) (M) are calculated where:

N = the number of molecules already in the arrangement

M = the number of bushes on the current sample being arranged

and C_F and C_R have already been described.

This gives a measure of how well a new molecule fits into the present arrangement. The correlation cutoff value entered in response to:

CORRELATION CUTOFF VALUE?

will be the minimum value either of the above values must have before a sample is added to the present arrangement. Entering a zero (0) causes none of the checking to be performed. This is another method the user may employ to add those samples to the current arrangement which fit best. The others may be added in a subsequent re-analysis (see below).

f) Starting the Arrangement in the Middle - once the Best Fit Histogram has been performed on the initial samples in the data set, saving that arrangement on the disk will cause those samples to remain in that orientation for subsequent retrievals. Because the user will want to perhaps perform several analyses on the same data set without repeating the Best Histogram Fit, he may respond "Y" to the question:

START ANALYSIS IN THE MIDDLE (Y,N)?

which will be followed by the question:

START WITH WHICH MOLECULE?

For any n which the user may respond with ($1 \leq n \leq k$ where k is the number of samples in the data set), the program will automatically form a model using the first n-1 samples in the data set in their current orientation.

Therefore if the user has previously saved some data upon which a Best Histogram Fit of 7 samples was performed, specifying 8 to the question above will cause the same Best Fit model to be used and start the subsequent arrangement with the eighth sample.

This feature also allows the user to take a final arrangement of his current data and determine if adding more samples to the data set will improve the present results. If a user has 75 arranged samples in a data file and wishes to add 50 more in an attempt to improve the results, he simply adds the 50 new samples using DNA1 and responds "N" to the question:

DO YOU WANT THE DATA PUT IN ORDER (Y,N)?

After saving this data to the disk he may now perform an arrangement of the new data into the old data by specifying that the arrangement start in the middle of the data set at the 76th sample.

g) Reanalysis - once all of the parameters have been chosen, the analysis begins. All of the sample numbers are displayed on the video monitor as they are arranged. Since it is possible that certain samples will not be included in the arrangement because they don't meet the requirements of the user defined Correlation Cutoff Value, those sample numbers are displayed with an asterisk (*) next to them. If, at the end of an analysis, certain samples have been left out of the arrangement, the user will be given a message indicating that fact and be allowed to re-analyze the data in an attempt to include those samples. The histogram coefficient, C_H, is printed at the end of each analysis to indicate how the arrangement is shaping up.

If the user does not wish to attempt to fit the remaining samples into the arrangement at this time, he responds "N" to the question:

RE-ANALYZE (Y,N)?

The user may wish to first generate a histogram of the current arrangement before performing a re-analysis. This may be accomplished by responding "Y" to the question:

WOULD YOU LIKE A HISTOGRAM (Y,N)?

and following the directions described under "HISTOGRAMS" below.

If the user wishes to perform the reanalysis, he responds "Y" to the above question. He will then be allowed to change the Correlation Cutoff Value before performing the reanalysis.

After performing any of a number of analyses on the data, the user may generate a histogram, following the directions below.

h) Miscellaneous Features - if the user wishes, he may terminate the analysis currently in progress by typing an "S". This action will terminate the arrangement option and return him to the point of choosing between the 3 major options provided by DNA2.

3) HISTOGRAMS:

The purpose of this option is to provide the user with a graphic representation of the current arrangement of his data either on the video monitor only or as a hardcopy on the line printer. The process of obtaining histograms is initiated by responding "Y" to the question:

WOULD YOU LIKE A HISTOGRAM (Y,N)?

Any number of histograms of the current data may be obtained using the options described below. After producing the desired histograms, statistics may be generated based on the current arrangement of the data.

Options Available When Producing Histograms:

a) Step Size - this option allows the user to choose the size of the intervals on the x-axis of the histogram. We typically choose 1% divisions, thus generating 100 intervals.

b) Multiplication Factor - this option has been included to accommodate a special property of some of our data sets. We frequently wish to compare two different data sets where one set consists of total DNA molecules and the other of many occurrences of only a fragment of that molecule. After performing the arrangement on both sets of data, we wish to compare the histogram of the fragment data to its corresponding area on the histogram of the total molecule by placing the printouts alongside of each other. This would not be possible unless we were able to "squeeze" the histograms of the fragment so that its size was equal to the corresponding area on the total molecule histogram. This is accomplished by multiplying the bush positions of all of the fragment molecules by a percentage of the length of the total molecule.

For example, if we have a situation as above, where our fragment is the lefthand 58% of the total molecule, we need only to multiply the fragment data by 0.58 to "squeeze" it down to the size of the corresponding region of the histogram of the total molecule. For an example of this option se the section "Sample Outputs From DNA2".

c) Minimum and Maximum Bush Number - this option serves the purpose of limiting the samples which will be present in the histogram. This allows the user to take his data arrangement and produce histograms of samples with 1 bush, 2 bushes etc. allowing him to assess the role played by certain classes of samples in the composite histogram.

d) Reverse - this option allows the user to reverse the orientation of all of the samples in the data set. When statistics are performed (see below), they reflect the current orientation of the samples. When comparing statistics in 2 different data sets, it will be necessary to have both data sets in the same orientation.

e) Smoothing - this is an option which has proven to be very useful to us in our work. Ordinarily, when a histogram is produced, the values on the y-axis reflect the number of samples with a bush in a given interval. However, to analyze the data in this form is difficult because of the "noise" which occurs due to random binding of the proteins to the DNA. In addition, as has been explained previously, many occurrences of a bush at a specific site when measured, will be found to form a normal distribution around a central site. Our smoothing function takes an occurrence of a bush in a given interval and makes it into a small normal distribution. So, for each occurrence of a bush in interval i of H(i), $1 \leq i \leq k$, where k is the number of intervals, we form a smoothing array T, which is initially set to zero, as follows:

$$T(i+1) = .3413 + T(i+1)$$

$$T(i-1) = .3413 + T(i-1)$$

$$T(i+2) = .1359 + T(i+2)$$

$$T(i-2) = .1359 + T(i-2)$$

After performing the calculations above for all intervals, we add the smoothing array, T, to the data array, H, as follows:

$$H(i) = H(i) + T(i) \qquad 1 \leq i \leq k$$

Examples of the effect of this option are given in the section "Sample Outputs From DNA2".

f) On-Screen Options - after the user has chosen whether or not he wants his histogram to be smoothed, the histogram will appear on the video monitor, with divisions on the x-axis of 5% and divisions on the y-axis of 1 sample each (includes smoothed array when that option is chosen). This will remain on the screen while the user exercises the options below:

I) Hardcopy - typing a "C" produces a hardcopy version of the histogram which is currently on the screen on the line printer. Information provided with the plot includes:

FILENAME

STEP SIZE

MULTIPLICATION FACTOR

RANGE OF BUSHES INCLUDED

HISTOGRAM COEFFICIENT (C_H)

These parameters serve to provide the user with the means of reproducing the plot at some other time.

II) Reverse Screen - typing an "R" will cause the entire plot to be reversed on the screen (from left to right), thus giving a mirror image of the original plot. This in combination with Option I allows the user to quickly obtain hardcopy plots of both orientations of one arrangement. It should be noted that this option reverses the screen image only and not the data itself. When statistics are performed (see below), they will reflect the current orientation of the data only.

III) Continue Analysis Which Is Currently In Progress - if the user has chosen not to re-analyze his data so that he might first obtain a histogram of the current arrangement of the data, typing an "A" will cause the re-analysis option to be given once again. Thus, the user might alternately run an analysis and generate a histogram to see how the arrangement is shaping up graphically.

IV) Begin Statistics - once the user has obtained all of the graphics he desires, typing the "RETURN" key causes the program to go into the statistics mode. The statistics allow the user to assess the probability that the arrangement he has generated accurately represents the real situation.

g) Statistics - three types of statistics are provided so that the user may assess his results. They are as follows:

Type 1 - Single Site Statistics - this allows the user to obtain statistics based on individual peaks in the histogram. The user inputs the range which a peak covers in response to the question:

```
INPUT RANGE (MIN,MAX)?
```

The statistics provided include:

The number of samples with a bush in the specified range

Mean measurement of the included bushes

Standard deviation

Variance

Entering "0,0" in response to the above question terminates Type I statistics.

Type II) Multiple Site Statistics - this option allows the user to determine whether a peak in the histogram is real or an accidental result of the arrangement process. Ranges are input for a number of peaks, specified in response to the question:

```
NUMBER OF SITES?
```

If any samples are found with bushes in all of the specified ranges, its identifying number is printed out.

This option is useful in the situation where the user would like to determine whether a given peak on a histogram is real. By inputting the range of that peak and also the range of the peak assumed to be real based on other results, it can be determined whether or not there are any samples with bushes in both the known peak and the undecided peak. If not, then the undecided peak may not be real. If a number of samples do exist, then it might be assumed that both peaks are real.

Type III) Distribution Statistics - this option allows the user to determine whether a wide peak is an accurate representation of the real situation. After inputting the range of a peak, a distribution of the number of bushes within the range will be generated of the form:

BUSHES IN RANGE	MOLECULES
1	7
2	3
etc.	

If any samples are present which have a number of bushes in the specified range, the peak might be as wide as it appears in the histogram.

```
SAMPLE OUTOUTS FROM DNA2:

(I) Correlations

RUN
----------------------------------------
              DNA ANALYSIS
----------------------------------------

FILENAME?testdata1                          We'll be analyzing some data
                                            which we've produced in our
----------------------------------------    lab as an example of how to
DATA DESCRIPTION:                           use DNA2

EXAMPLE ILLUSTRATING DNA2

NO. MOLECULES= 86

----------------------------------------
                                            First we'll use the correlations
ARRANGEMENT, HISTOGRAM OR CORRELATIONS      option to find out which
(A,H,C)?c                                   samples are the most similar
----------------------------------------
INTERVAL SIZE (1-100)?1
                                            We'll limit the calculations to
MINIMUM AND MAXIMUM BUSH NUMBER?4,10        those samples with 4-10 bushes

MINIMUM CORRELATION FOR PRINTOUT?0          We'll have all correlations
                                            printed out


-----------------------------------------------------------------------------
  MOL/SITES         MOL/SITES          TOTAL                PER SITE
-----------------------------------------------------------------------------
                                    F     R     D        F     R     D
                                   -----------------    -----------------

   5769/4            5741/4         2     2     0        .5    .5    0
   5769/4            5751/4         2     .14   1.86     .5    .03   .46
   5769/4            5778/4         .14   0     .14      .03   0     .03
   5769/4            8809/4         0     2     2        0     .5    .5
   5769/4            8810/4         .27   0     .27      .06   0     .06
   5741/4            5751/4         .14   2.34  2.2      .03   .58   .55
   5741/4            5778/4         .34   2.14  1.8      .08   .53   .45
   5741/4            8809/4         .34   0     .34      .08   0     .08
   5741/4            8810/4         .14   2     1.86     .03   .5    .46
   5751/4            5778/4         .48   0     .48      .12   0     .12
   5751/4            8809/4         0     .14   .14      0     .03   .03
   5751/4            8810/4         .34   0     .34      .08   0     .08
   5778/4            8809/4         0     .34   .34      0     .08   .08
   5778/4            8810/4         2     0     2        .5    0     .5
   8809/4            8810/4         .34   .48   .14      .08   .12   .03
```

```
------------------------------------------------------------

-----------------------------------
WOULD YOU LIKE CERTAIN SAMPLES PRINTED
OUT (Y,N)?y
-----------------------------------
```

We'll print out all of the samples which have been shown to be similar

```
INPUT SAMPLE NUMBERS:  (TYPE '0' TO STOP)
?5769
?5741
?5751
?5778
?8809
?8810
?0
```

Terminates sample # input

```
-----------------------------------
5769
37 45 72 87  /  12 27 54 62
```

Samples #5769 and #5741 have C_F and C_R both equal to 2

```
-----------------------------------
5741
12 22 45 81  /  18 54 77 87
```

They are probably not a good choice for the model histogram.

```
-----------------------------------
5751
5 21 45 56  /  43 54 78 94
```

Samples #5741 and #5751 match in one direction only

```
-----------------------------------
5778
52 69 77 82  /  17 22 30 47
```

Sample #5778 is similar to #8810 as well as #5741 and #5769

```
-----------------------------------
8809
34 62 76 91  /  8 23 37 65
-----------------------------------
8810
14 64 77 99  /  0 22 35 85
```

It seems that these 5 samples are quite similar so we'll use them in the Best Histogram Fit when we begin arranging the samples

```
-----------------------------------
WOULD YOU LIKE CERTAIN MOLECULES PRINTED
OUT (Y,N)?n
-----------------------------------
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
-----------------------------------
FUNCTION?4

WOULD YOU LIKE TO SAVE THE CURRENT
  DATA (Y,N)?n
```

If the samples above had been unsatisfactory we could have rerun the correlations to include the samples with three bushes. For the sake of brevity, that won't be illustrated here.

(II) Arrangement

```
RUN
------------------------------------
            DNA ANALYSIS
------------------------------------

FILENAME?testdata1

------------------------------------

DATA DESCRIPTION:

EXAMPLE ILLUSTRATING DNA2

NO. MOLECULES= 86

------------------------------------

ARRANGEMENT, HISTOGRAM OR CORRELATIONS
(A,H,C)?a                                  We're ready to start the
                                           arrangement process

DO YOU WANT TO PUT CERTAIN SAMPLES
   FIRST (Y,N)?y

HOW MANY SAMPLES (1-10)?5                  We'll move the 5 similar samples
                                           which we've chosen to the top
INPUT SAMPLE NUMBERS BELOW:                of the data set

SAMPLE #1?57699                            Error in entering sample #

SAMPLE # 57699 NOT FOUND

SAMPLE #1?5769

SAMPLE #2?5751

SAMPLE #3?8809

SAMPLE #4?5741

SAMPLE #5?5769

SAMPLE # 5769 HAS ALREADY BEEN MOVED       Accidentally repeated #5769

SAMPLE #5?5778

INTERVAL SIZE (1-100)?1                    We'll choose 1% intervals

DO YOU WANT TO DO A BEST HISTOGRAM FIT
```

```
   ON THE FIRST FEW SAMPLES (Y,N)?y

HOW MANY SAMPLES (1-10)?5

WOULD YOU LIKE THE BEST-FIT ARRANGEMENT
  PRINTED OUT (Y,N)?y
----------------------------------------
SAMPLE #5769
           37  45          72  87
----------------------------------------
SAMPLE #5751
5      21      45  56
----------------------------------------
SAMPLE #8809
8      23  37          65
----------------------------------------
SAMPLE #5741
   12  22      45              81
----------------------------------------
SAMPLE #5778
   17  22  30  45
----------------------------------------
```

Begin Best Histogram Fit analysis on the first five samples in the data set

Once completed, we can print out the results with the bush positions shown relative to each other

```
MINIMUM AND MAXIMUM BUSH NUMBER?1,10

CORRELATION CUTOFF VALUE?.25

THE ANALYSIS INCLUDES THE FOLLOWING
   SAMPLES:

5769           5751           8809
5741           5778           8810
5774*          5739*          5753*
5762*          5765*          5766
5756*          5757*          5763*
5738*          5731*          5732*
5733*          5734*          5749
5754           5758*          5760*
5764*          5775*          5777*
8814*          8788*          8802*
8803*          9873*          9864*
5745*          5772           5728*
5735           5736*          5740*
5744*          5748*          5755*
5759*          5767*          5768*
5770*          5776*          8813*
8875*          8794*          8795*
9871           9877*          9879*
9882*          9864.2*        9935*
9942*          5729*          5737*
5742*          5746*          5747*
```

Samples marked with an asterisk (*) did not meet the correlation cutoff requirements

```
5774.2*        5684*        8871*
8824*          8826*        9793*
8796*          8801*        8804*
9865*          9868*        9874*
9878*          9861*        9862*
9625           9927*        9931
9937*          9945*        9946
9949           9952*

HISTOGRAM COEFFICIENT= 2.63
```

Current C_H value

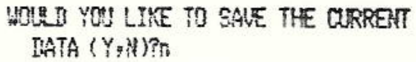
```
SOME SAMPLES WERE LEFT OUT OF THIS
ARRANGEMENT.  RE-ANALYZE (Y,N)?n

WOULD YOU LIKE A HISTOGRAM (Y,N)?n
```

We'll save this arrangement
to preserve the Best Histogram Fit
which we currently have

```
-----------------------------------

YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
-----------------------------------
FUNCTION?3
-----------------------------------

INPUT DATA DESCRIPTION (ONE LINE)

?testdata with b.h.f. of 5

FILENAME?testdata2
```

New filename

```
DATA HAS BEEN SAVED

-----------------------------------
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
-----------------------------------
FUNCTION?4
```



```
WOULD YOU LIKE TO SAVE THE CURRENT
  DATA (Y,N)?n
```

```
RUN
----------------------------------------
            DNA ANALYSIS
----------------------------------------
```

Now we'll do some analysis on the stored data

```
FILENAME?testdata2

----------------------------------------

DATA DESCRIPTION:

TESTDATA WITH B.H.F. OF 5

NO. MOLECULES= 86

----------------------------------------

ARRANGEMENT, HISTOGRAM OR CORRELATIONS
(A,H,C)?a
```

We'll be arranging the data

```
DO YOU WANT TO PUT CERTAIN SAMPLES
   FIRST (Y,N)?n
```

We already have our Best Fit samples in place

```
INTERVAL SIZE (1-100)?1

DO YOU WANT TO DO A BEST HISTOGRAM FIT
   ON THE FIRST FEW SAMPLES (Y,N)?n
```

This has already been done

```
MINIMUM AND MAXIMUM BUSH NUMBER?1,10

CORRELATION CUTOFF VALUE?.2

START ANALYSIS IN THE MIDDLE (Y,N)?y

START WITH WHICH SAMPLE?6
```

We'll start with the sixth sample, thereby causing the first 5 samples to be used as the model

```
THE ANALYSIS INCLUDES THE FOLLOWING
   SAMPLES:

5769          5751          8809
5741          5778          8810
5774*         5739          5753*
5762*         5765*         5766
5756*         5757*         6763*
5738*         5731*         5732*
5733*         5734*         5749
5754          5758*         5760*
5764*         5775*         5777*
8814*         8788*         8802*
8803*         9873          9864*
5745          5772          5728*
```

```
5735          5736*         5740*
5744*         5748*         5755*
5759*         5767*         5768*
5770          5776*         8813*
8875*         8794*         8795*
9871          9877*         9879*
9882*         9864.2*       9935*
9942*         5729*         5737*
5742*         5746*         5747*
5774.2*       5684*         8871*
8824*         8826*         9793*
8796          8801*         8804*
9865*         9868*         9874*
9878          9861*         9862*
9625*         9927*         9931*
9937*         9945*         9946*
9949          9952

HISTOGRAM COEFFICIENT= 3.04

SOME SAMPLES WERE LEFT OUT OF THIS
ARRANGEMENT.  RE-ANALYZE (Y,N)?y

CHANGE CORRELATION CUTOFF VALUE (Y,N)y

NEW CUTOFF VALUE?.1
```

We'll lower the correlation
cutoff value to ease the
requirements for samples now that
we have a model with more samples

```
REANALYSIS:

THE ANALYSIS INCLUDES THE FOLLOWING
   SAMPLES:

5769          5751          8809
5741          5778          8810
5774          5739          5753*
5762          5765*         5766
5756*         5757*         5763
5738*         5731          5732*
5733*         5734*         5749
5754          5758*         5760*
5764          5775*         5777
8814*         8788*         8802*
8803          9873          9864*
5745          5772          5728*
5735          5736          5740*
5744*         5748*         5755
5759          5767          5768*
5770          5776          8813
8875*         8794          8795*
9871          9877*         9879*
9882*         9864.2*       9935
```

```
9942*           5729*           5737*
5742            5746            5747*
5774.2          5684*           8871*
8824*           8826*           9793*
8796            8801*           8804
9865            9868*           9874*
9878            9861*           9862*
9625*           9927*           9931*
9937*           9945*           9946*
9949            9952

HISTOGRAM COEFFICIENT= 3.8

SOME SAMPLES WERE LEFT OUT OF THIS
ARRANGEMENT.  RE-ANALYZE (Y,N)?y

CHANGE CORRELATION CUTOFF VALUE (Y,N)?y

NEW CUTOFF VALUE?0
```

Now we'll allow all of the remaining samples to be entered

```
REANALYSIS:

THE ANALYSIS INCLUDES THE FOLLOWING
   SAMPLES:

5769            5751            8809
5741            5778            8810
5774            5739            5753
5762            5765            5766
5756            5757            5763
5738            5731            5732
5733            5734            5749
5754            5758            5760
5764            5775            5777
8814            8799            8802
8803            9873            9864
5745            5772            5728
5735            5736            5740
5744            5748            5755
5759            5767            5768
5770            5776            8813
8875            8794            8795
9871            9877            9879
9882            9864.2          9935
9942            5729            5737
5742            5746            5747
5774.2          5684            8971
8824            8826            9793
8796            8801            8804
9865            9868            9874
9878            9861            9862
```

```
9625              9927              9931
9937              9945              9946
9949              9952

HISTOGRAM COEFFICIENT= 3.87

WOULD YOU LIKE A HISTOGRAM (Y,N)?n
```

We'll save the data before obtaining a histogram

```
______________________________________

YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
______________________________________
FUNCTION?3
______________________________________

INPUT DATA DESCRIPTION (ONE LINE)

?arranged data of testdata2 - 3/26

FILENAME?testdata2

TESTDATA2 ALREADY IN CATALOG

OVERWRITE (Y,N)?n
```

We almost wrote over the original data
We'll put this arrangement in another file

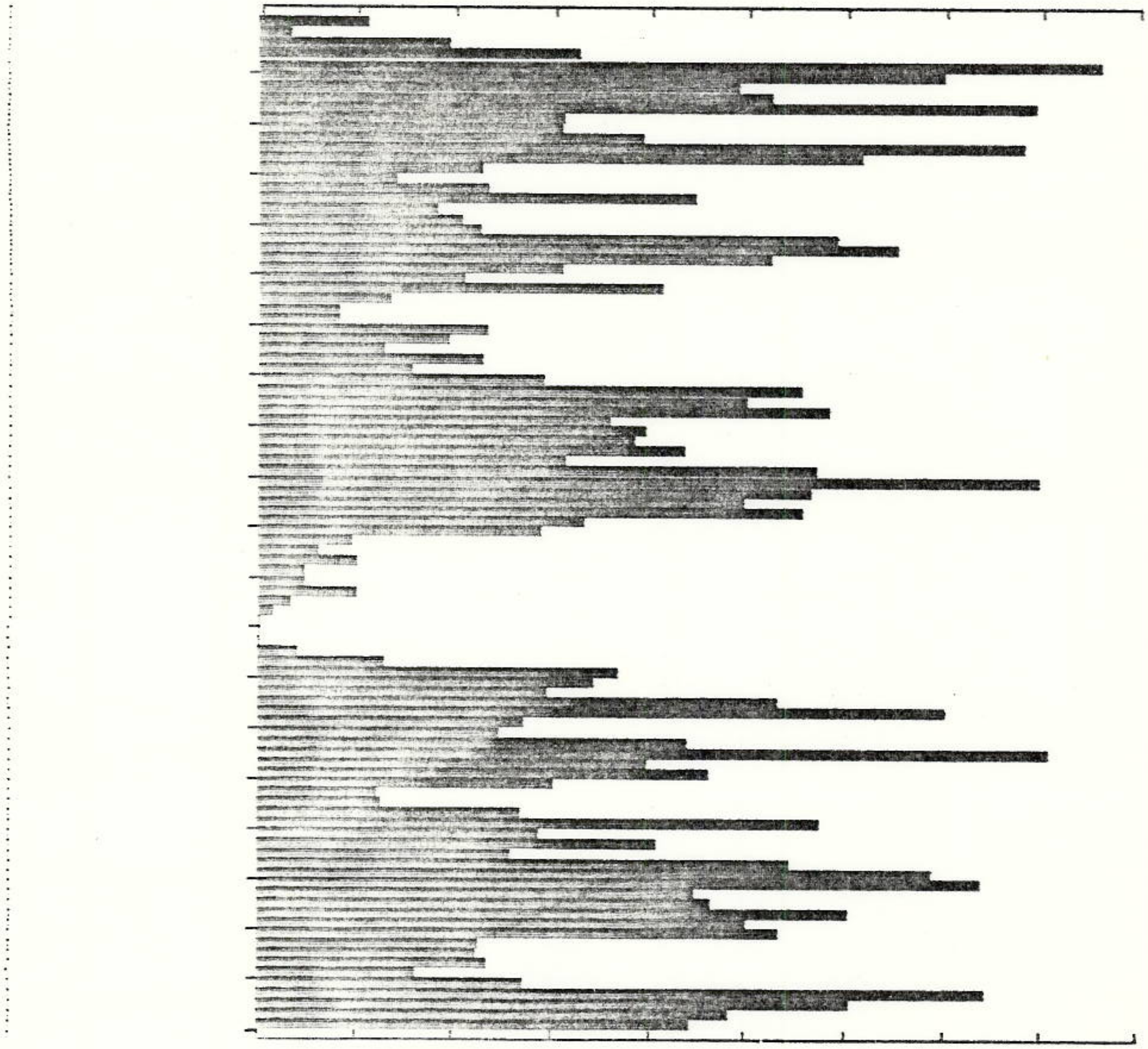
```
FILENAME?testdata3

DATA HAS BEEN SAVED

______________________________________
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
______________________________________
FUNCTION?4

WOULD YOU LIKE TO SAVE THE CURRENT
  DATA (Y,N)?n
```

(III) Histograms

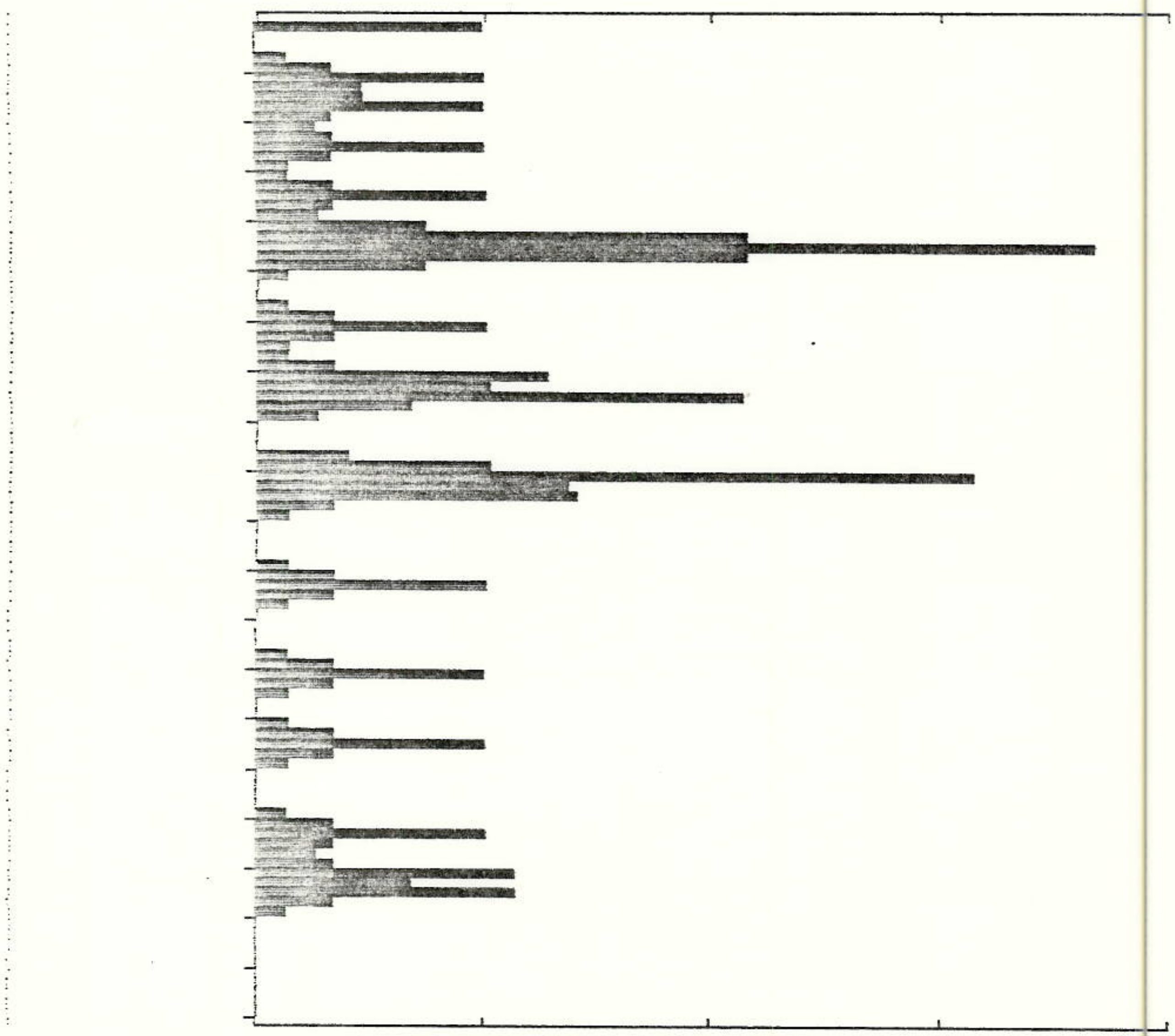
```
RUN
-----------------------------------
           DNA ANALYSIS
-----------------------------------
FILENAME?testdata3
-----------------------------------
DATA DESCRIPTION:
ARRANGED DATA OF TESTDATA2 - 3/26
NO. MOLECULES= 86
-----------------------------------
ARRANGEMENT, HISTOGRAM OR CORRELATIONS
(A,H,C)?h                                   Now we'll obtain some histograms
                                            of our data arrangement
-----------------------------------
HISTOGRAM STEP SIZE (1-100)?1               Interval size of 1%
MULTIPLICATION FACTOR?1
MINIMUM AND MAXIMUM BUSH NUMBER?1,10        Include all samples
REVERSE (Y,N)?n
SMOOTHING (Y,N)?y                           At this point the histogram is
                                            displayed on the video monitor
WOULD YOU LIKE STATISTICS (Y,N)?n           Typing 'C' prints it on the line
                                            printer (See USING DNA2)
-----------------------------------
HISTOGRAM STEP SIZE (1-100)?1
MULTIPLICATION FACTOR?1                     Now we'll produce 2 more
                                            histograms to see the role
MINIMUM AND MAXIMUM BUSH NUMBER?4,4         played by samples with 4 bushes
                                            and 3 bushes (See next pages
REVERSE (Y,N)?n                             for actual histograms)
SMOOTHING (Y,N)?y
WOULD YOU LIKE STATISTICS (Y,N)?n
```

```
----------------------------------------

HISTOGRAM STEP SIZE (1-100)?1

MULTIPLICATION FACTOR?1

MINIMUM AND MAXIMUM BUSH NUMBER?3,3

REVERSE (Y,N)?n

SMOOTHING (Y,N)?y

WOULD YOU LIKE STATISTICS (Y,N)?n

----------------------------------------

HISTOGRAM STEP SIZE (1-100)?0              Terminates histogram generation

----------------------------------------
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
----------------------------------------
FUNCTION?4

WOULD YOU LIKE TO SAVE THE CURRENT
   DATA (Y,N)?n
```

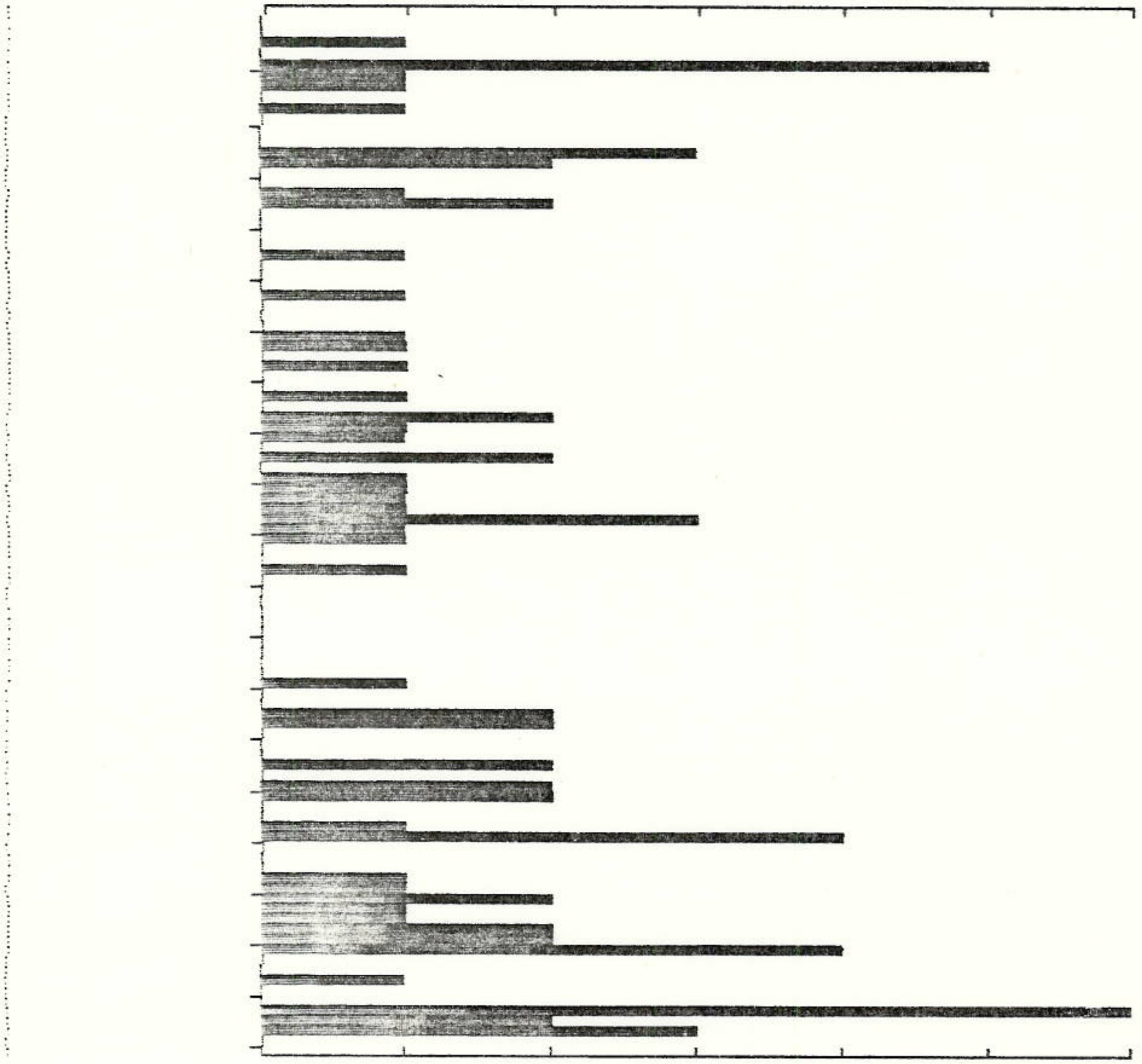
```
--------------------------------------------
FILENAME: TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 4.89466912
--------------------------------------------
```



```
-----------------------------------------
FILENAME:  TESTDATA3
BYTESIZE= 1
MULTIPLICATION FACTOR= 1
PACKED= 4
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 1.37152806
-----------------------------------------
```



```
---------------------------------------
FILENAME:  TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 3
SMOOTHING: N
HISTOGRAM COEFFICIENT= 2.45679012
---------------------------------------
```



Below are some sample outputs illustrating the SMOOTHING and REVERSE (on screen) options.

```
RUN
-----------------------------------
            DNA ANALYSIS
-----------------------------------

FILENAME?testdata3

-----------------------------------

DATA DESCRIPTION:

ARRANGED DATA OF TESTDATA2 - 3/26

NO. MOLECULES= 86

-----------------------------------

ARRANGEMENT, HISTOGRAM OR CORRELATIONS
(A,H,C)?h

-----------------------------------

HISTOGRAM STEP SIZE (1-100)?1

MULTIPLICATION FACTOR?1

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?n
```

First we'll produce an unsmoothed and unreversed histogram

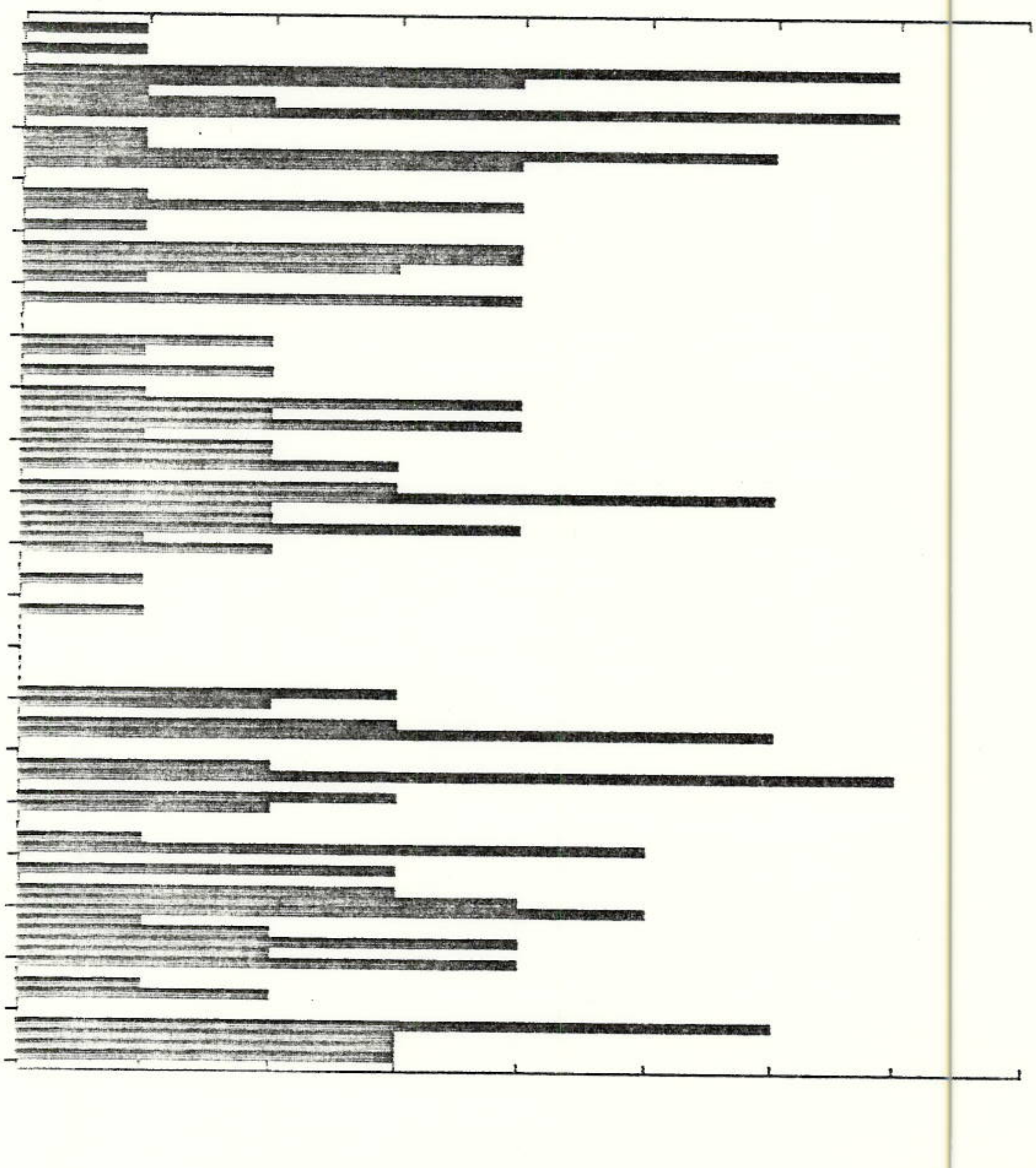
```
SMOOTHING (Y,N)?n

WOULD YOU LIKE STATISTICS (Y,N)?n

-----------------------------------

HISTOGRAM STEP SIZE (1,100)?1

MULTIPLICATION FACTOR?1

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?n

SMOOTHING (Y,N)y
```

We'll smooth this one to show how smoothing affects the histogram

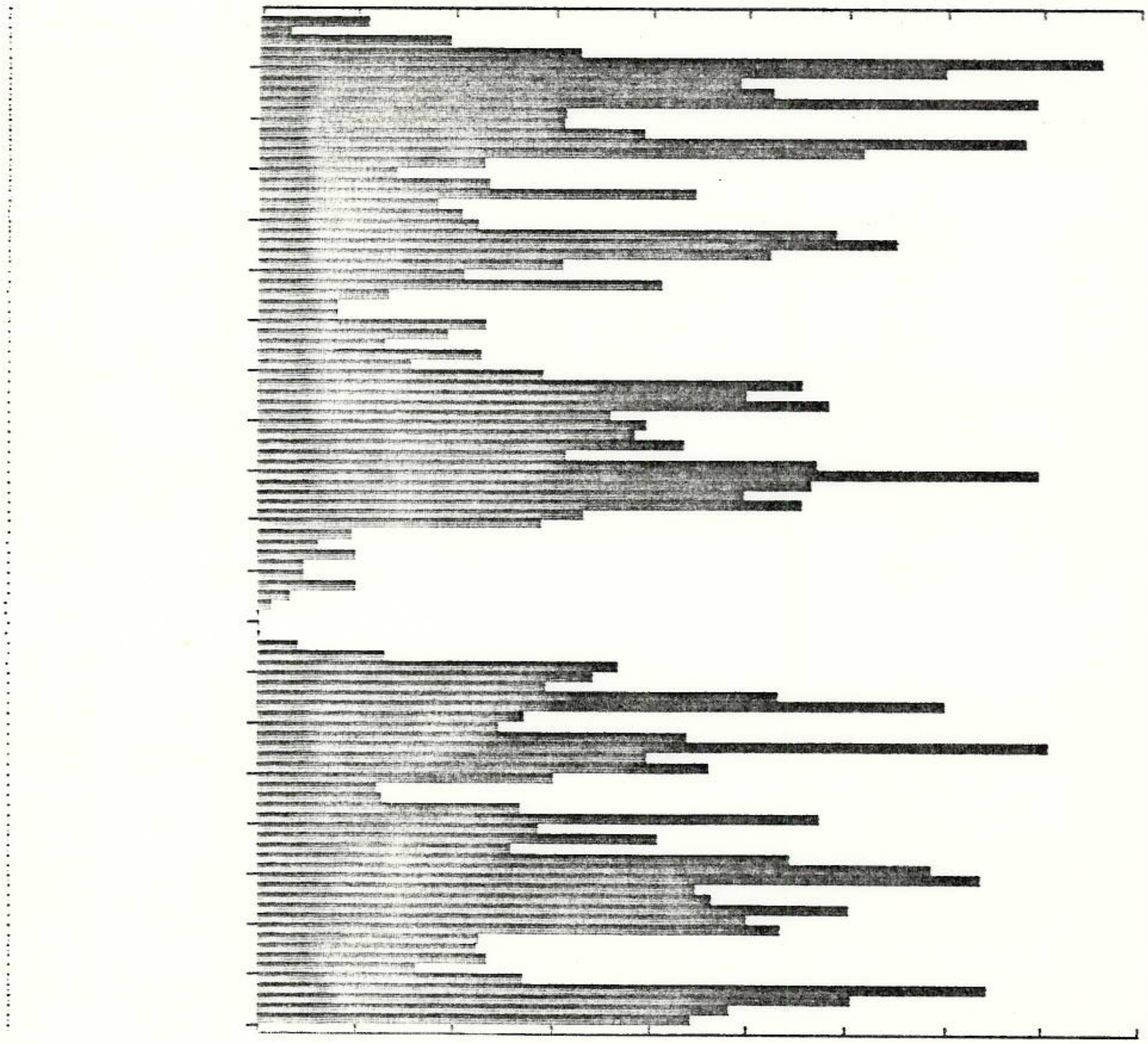
```
WOULD YOU LIKE STATISTICS (Y,N)?n
```

```
------------------------------------

HISTOGRAM STEP SIZE (1-100)?1

MULTIPLICATION FACTOR?1

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?y                              We'll reverse this one to
                                             illustrate that option
SMOOTHING (Y,N)?y

WOULD YOU LIKE STATISTICS (Y,N)?n

------------------------------------

HISTOGRAM STEP SIZE (1-100)?0                Terminates histogram generation

------------------------------------
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
------------------------------------
FUNCTION?4

WOULD YOU LIKE TO SAVE THE CURRENT
   DATA (Y,N)?n
```

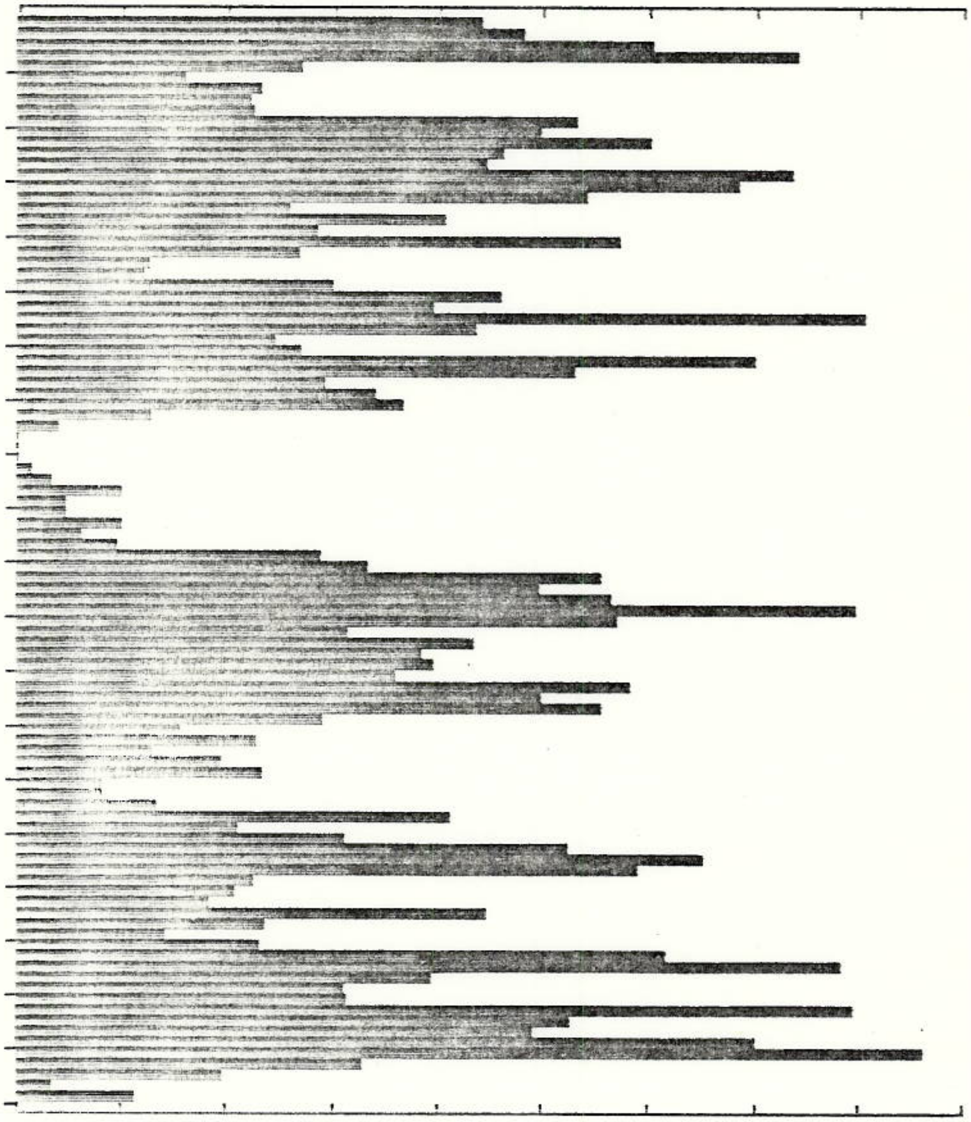
```
----------------------------------------
FILENAME: TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: N
HISTOGRAM COEFFICIENT= 3.87431694
----------------------------------------
```



```
----------------------------------------
FILENAME:  TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 4.89466912
----------------------------------------
```



```
----------------------------------------
FILENAME:  TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
PHONES= 1 -10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 4.09446912
----------------------------------------
```

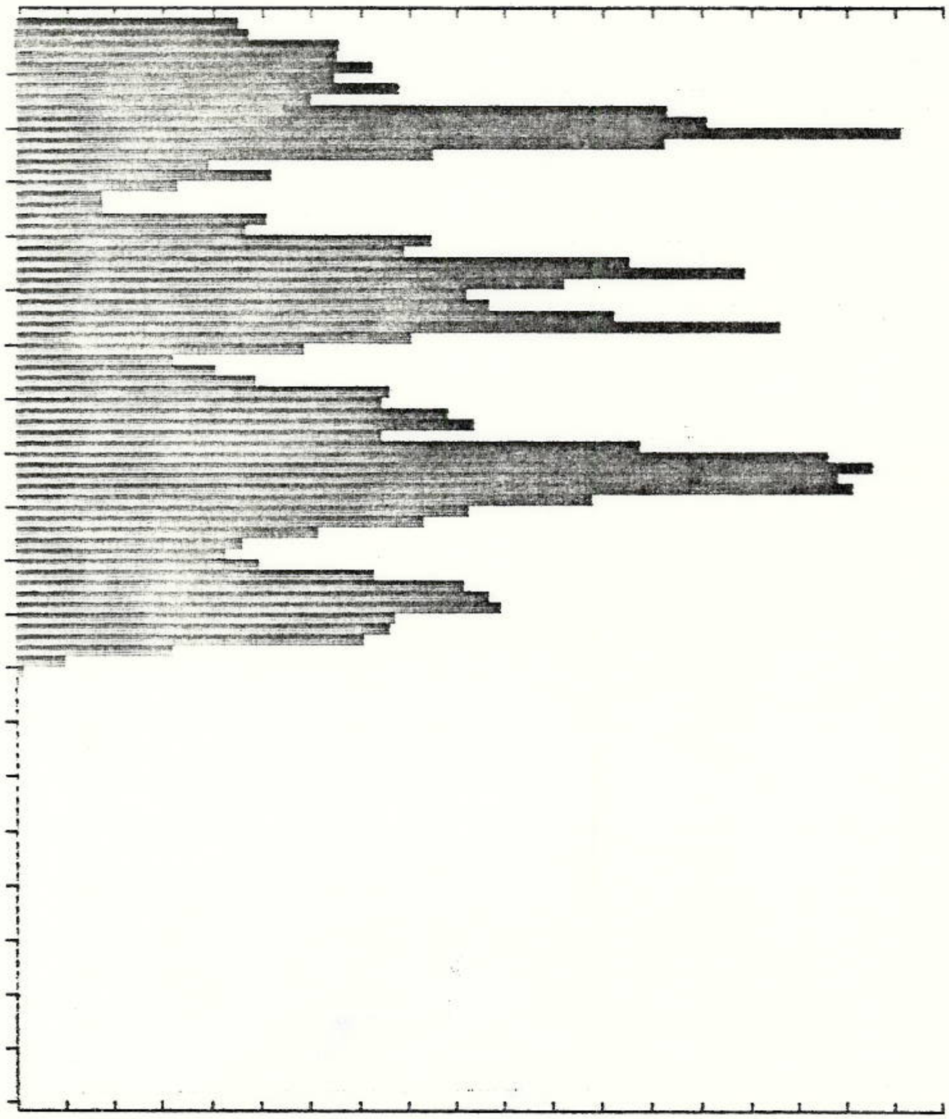


These examples illustrate the use of the MULTIPLICATION FACTOR option. The data illustrated here consists of fragments which are known to be the lefthand 58.5% of the total molecule data in TESTDATA3. Stepsizes of 1% and 2% are shown.

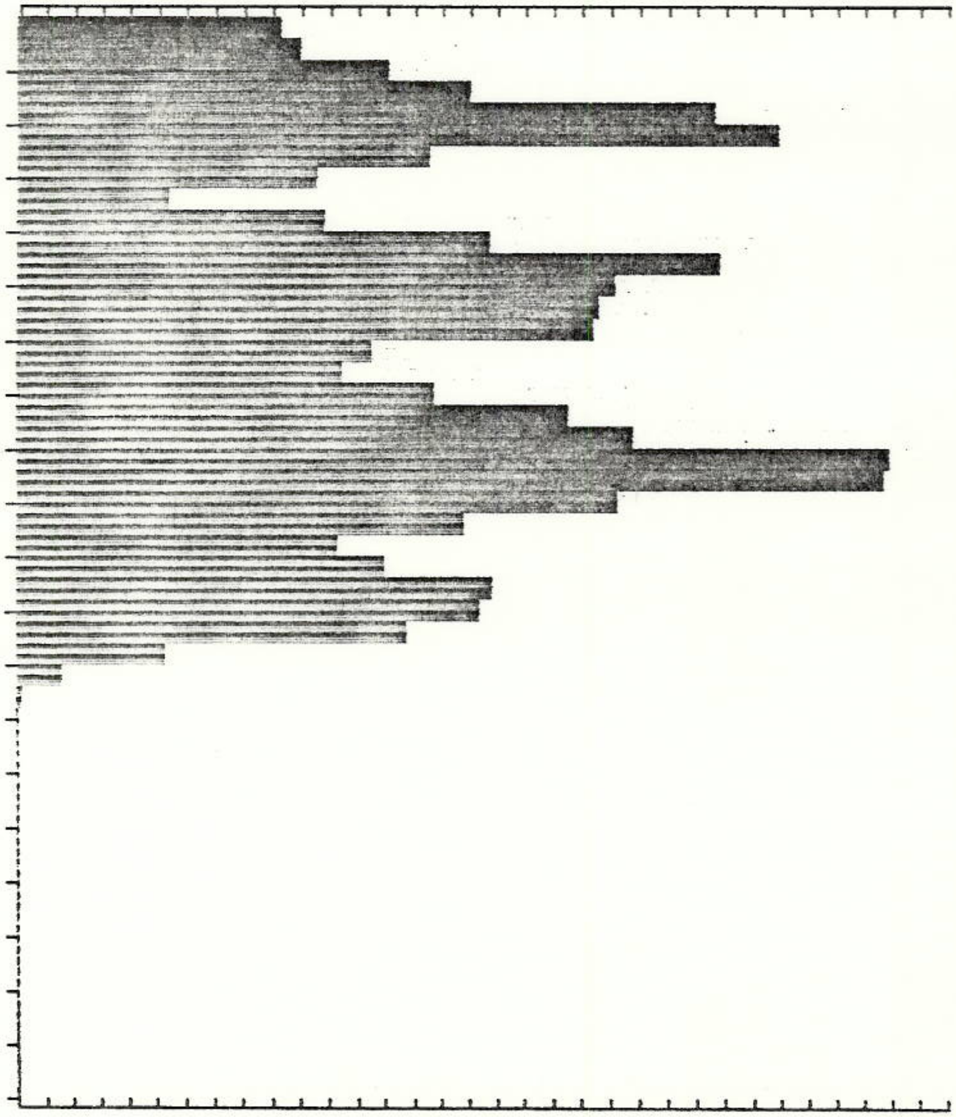
```
RUN
--------------------------------------
           DNA ANALYSIS
--------------------------------------

FILENAME?arrwspfraga first 11/10

--------------------------------------

DATA DESCRIPTION:

FIRST ARRANGEMENT OF FRAGA 2MIN BUSH

NO. MOLECULES= 83

--------------------------------------

ARRANGEMENT, HISTOGRAMS OR CORRELATIONS
(A,H,C)?h

--------------------------------------

HISTOGRAM STEP SIZE (1-100)?1

MULTIPLICATION FACTOR?.585

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?n

SMOOTHING (Y,N)?y

WOULD YOU LIKE STATISTICS (Y,N)?n

--------------------------------------

HISTOGRAM STEP SIZE (1-100)?2

MULTIPLICATION FACTOR?..585

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?n

SMOOTHING (Y,N)?y
```

```
WOULD YOU LIKE STATISTICS (Y,N)?n

------------------------------------

HISTOGRAM STEP SIZE (1-100)?0

------------------------------------
YOU MAY NOW:

     1) CONTINUE WITH CURRENT DATA

     2) READ IN SOME OTHER DATA

     3) SAVE CURRENT ARRANGEMENT

     4) TERMINATE PROCESSING
------------------------------------

FUNCTION?4

WOULD YOU LIKE TO SAVE THE CURRENT
   DATA (Y,N)?n
```

```
----------------------------------------
FILENAME: ARRWSPERASA FIRST 11/10
STEPSIZE= 1
MULTIPLICATION FACTOR= .585
PULSES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 10.5471675
----------------------------------------
```



```
------------------------------------------
FILENAME:  ARRWGPFRAGA FIRST 11/10
STEPSIZE= 2
MULTIPLICATION FACTOR= .595
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 19.2289365
------------------------------------------
```



(IV) Statistics

The statistics options are illustrated here using TESTDATA3 as an example. The histogram upon which the statistics are performed is given after the following examples.

```
RUN
-----------------------------------
           DNA ANALYSIS
-----------------------------------

FILENAME?testdata3

-----------------------------------

DATA DESCRIPTION:

ARRANGED DATA OF TESTDATA2 - 3/26

NO. MOLECULES= 86

-----------------------------------
ARRANGEMENT, HISTOGRAM OR CORRELATIONS

(A,H,C)?h

-----------------------------------

HISTOGRAM STEP SIZE (1-100)?1

MULTIPLICATION FACTOR?1

MINIMUM AND MAXIMUM BUSH NUMBER?1,10

REVERSE (Y,N)?n

SMOOTHING (Y,N)?y

WOULD YOU LIKE STATISTICS (Y,N)?y
-----------------------------------
  1) PRINT STATISTICS ON BUSHES WITHIN
     A SINGLE RANGE
  2) PRINT ALL MOLECULES WITH A BUSH
     IN EACH OF A NUMBER OF RANGES
  3) PRINT A DISTRIBUTION OF THE NUMBER
     OF BUSHES IN A GIVEN RANGE FOR ALL
     MOLECULES
-----------------------------------
```

```
WHICH TYPE (1-3)?1
```

Choose Type I statistics

```
-----------------------------------
FILENAME:  TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 5.00131727
-----------------------------------
```

Since these statistics are produced on the video monitor and on the line printer, this information about the data is given to identify the output

```
INPUT RANGE (MIN,MAX)?3,10
```

The statistics given will be a reflection of all samples with a bush in the range 3-10%

```
-----------------------------------

RANGE= 3 - 10

NUMBER OF MOLECULES= 20

MEAN MEASUREMENT= 6.06833229


STD. DEV.= 1.72730879

VARIANCE= 2.98359564

-----------------------------------
INPUT RANGE?12,21

-----------------------------------

RANGE= 12 - 21

NUMBER OF MOLECULES= 19

MEAN MEASUREMENT= 16.1302596

STD. DEV.= 1.94351152

VARIANCE= 3.77723704

-----------------------------------
INPUT RANGE?0,0
```

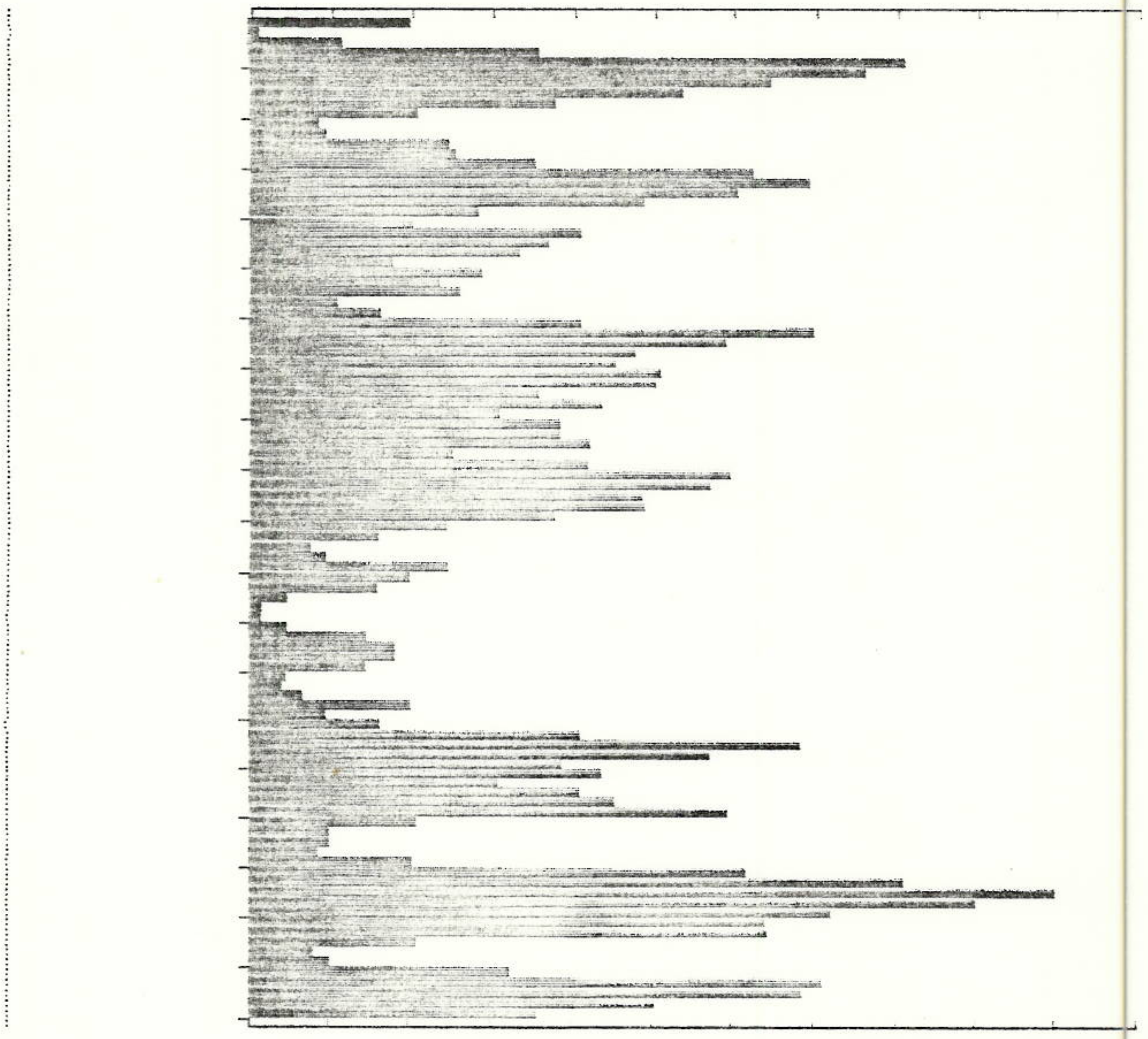
Terminate Type I statistics

```
-----------------------------------
WOULD YOU LIKE STATISTICS (Y,N)?y
-----------------------------------
  1) PRINT STATISTICS ON BUSHES WITHIN
     A SINGLE RANGE
  2) PRINT ALL MOLECULES WITH A BUSH
     IN EACH OF A NUMBER OF RANGES
  3) PRINT A DISTRIBUTION OF THE NUMBER
     OF BUSHES IN A GIVEN RANGE FOR ALL
     MOLECULES
```

```
-----------------------------------

WHICH TYPE (1-3)?2                        Choose Type II statistics
-----------------------------------
FILENAME: TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 5.00131727
-----------------------------------

NUMBER OF SITES?2                         We'll look at 2 peaks

RANGE 1?3,10

RANGE 2?12,21

-----------------------------------
RANGE #1  3-10
RANGE #2  12-21

5753          5766          8803          These three samples each have a
                                          bush in the ranges specified
-----------------------------------       above
NUMBER OF SITES?0                         Terminate Type II statistics

-----------------------------------
WOULD YOU LIKE STATISTICS (Y,N)?y
-----------------------------------
  1) PRINT STATISTICS ON BUSHES WITHIN
     A SINGLE RANGE
  2) PRINT ALL MOLECULES WITH A BUSH
     IN EACH OF A NUMBER OF RANGES
  3) PRINT A DISTRIBUTION OF THE NUMBER
     OF BUSHES IN A GIVEN RANGE FOR ALL
     MOLECULES
-----------------------------------

WHICH TYPE (1-3)?3                        Choose Type III statistics
-----------------------------------
FILENAME: TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 5.00131727
-----------------------------------

INPUT RANGE?83,93                         Input range of single peak
-----------------------------------
```

```
RANGE:  83 - 93

BUSHES IN RANGE     # MOLECULES
                                          28 samples have one bush each
      1                  28               in the specified range but only
      2                  1                1 has more than 1

-----------------------------------

INPUT RANGE?0,0                           Terminate Type III statistics

-----------------------------------

WOULD YOU LIKE STATISTICS (Y,N)?n         Terminate statistics option
-----------------------------------

HISTOGRAM STEP SIZE (1-100)?0             Terminate histogram option

-----------------------------------

YOU MAY NOW:

    1) CONTINUE WITH CURRENT DATA

    2) READ IN SOME OTHER DATA

    3) SAVE CURRENT ARRANGEMENT
    4) TERMINATE PROCESSING
-----------------------------------

FUNCTION?4                                Terminate processing

WOULD YOU LIKE TO SAVE THE CURRENT
  DATA (Y,N)?n
```

```
-----------------------------------------------
FILENAME:  TESTDATA3
STEPSIZE= 1
MULTIPLICATION FACTOR= 1
BUSHES= 1-10
SMOOTHING: Y
HISTOGRAM COEFFICIENT= 5.00131727
-----------------------------------------------
```



DISCUSSION:

It has been our experience that the system we have developed allows the user to perform an analysis, which previously took several weeks, in several hours instead. Since the research in this lab involves running many sets of data gathered by varying the conditions under which the proteins are allowed to bind to the DNA, this system has saved the users of this system much time when analyzing their data.

The most time consuming and most important portion of the analysis is the arrangement process. The time required here (assuming a correlation cutoff value of zero) is O(n). However, since it is possible that choosing a specific correlation cutoff value may cause no samples to be added to the arrangement, the process may actually be greater than O(n). Therefore, it is difficult, if not impossible, to derive a run time for n samples.

One method which deserves more investigation is one of a class of algorithms known as Greedy algorithms. This method works in stages by considering one input at a time (in this case, one of n samples). The goal is to find a feasible solution which either minimizes or maximizes a given objective function (in this case, maximizing C_H). The samples would be chosen from a data set which had previously been ordered by applying some selection function on them. Upon selection for entrance into a developing arrangement, a sample would have the objective function applied to it to see whether this sample should be We approximate this by calculating $|C_F - C_R|$ and comparing it to the user defined correlation cutoff value for each sample selected. Perhaps some function could be derived which orders all of the samples in the data set. This would give us a definable run time which would include applying the selection function, ordering the samples based on the selection function and performing the arrangement using the objective function.

I have attempted to keep the programs in this system general to allow other users with similar data to use them for their work without requiring a large amount of alteration. I hope I have succeeded in this goal.

```
1  REM ************************************************************************************
2  REM *     DNA1  -  DATA BASE PROGRAM FOR DNA DATA      -     LAST REVISED: 12/27/78       *
3  REM *     SCOTT MC COURT                               -     DEPT. OF COMPUTER SCIENCE    *
4  REM ************************************************************************************
100  DIM A(1000),B%(201),I(201),S(201),D(1000),C%(201),B(201),E(40),S1(201),I1(201),L(25)
101  ONERR  GOTO 6000
105  SPEED= 255: PRINT "NOMONI,O,C": HOME
130  GOSUB 9999: PRINT  TAB( 17);"DNA1": GOSUB 9999: PRINT
137  PRINT "PROGRAM FOR DATA ENTRY, RETRIEVAL AND": PRINT "MODIFICATION": PRINT : PRINT "NOTE
     : EACH FILE IS LIMITED TO 200": PRINT "       SAMPLES": PRINT
138  GOSUB 9999: PRINT : PRINT "NEW OR OLD DATA (NEW,OLD)";: INPUT A$: IF A$ = "NEW" THEN 145

139  IF A$ <  > "OLD" THEN 138
140  PRINT : GOSUB 3000: PRINT : GOTO 10010
141  REM *
142  REM *     ROUTINE TO ADD NEW DATA                                                  *
143  REM *
145  PRINT : GOSUB 9999: PRINT  TAB( 8);"***** INSTRUCTIONS *****": GOSUB 9999: PRINT "TO TER
     MINATE DATA ENTRY, ENTER 0 FOR": PRINT "SAMPLE #": PRINT
146  PRINT "TO START NEXT SAMPLE, ENTER 0,0": PRINT : PRINT "TO DELETE CURRENT SAMPLE, ENTER
     999,999": PRINT
150 X = 1:M = 0:B%(1) = 1: GOSUB 9999:F$ = ""
157  PRINT : PRINT "INPUT DATA DESCRIPTION (ONE LINE ONLY)": PRINT : INPUT D$: IF  LEN (D$) >
     40 THEN  PRINT : PRINT "** TOO LONG **": GOTO 157
160  FOR Y = 1 TO  LEN (D$):E(Y) =  ASC ( MID$ (D$,Y,1)): NEXT Y:E(0) =  LEN (D$)
165  PRINT : GOSUB 9999: PRINT "BEGIN DATA ENTRY:"
176  IF X = 201 THEN  PRINT "CAN'T ACCEPT ANY MORE DATA":I(X) =  - 1: GOTO 410
180  GOSUB 9999: PRINT : PRINT "SAMPLE #";: INPUT I(X): PRINT
181  FOR Q = 1 TO X - 1: IF I(Q) <  > I(X) OR X = 1 THEN  NEXT Q: GOTO 183
182  PRINT : PRINT "SAMPLE #";I(X);" ALREADY IN FILE.": PRINT : PRINT "RE-ENTER": PRINT : GOTO
     180
183 O8 = M: IF I(X) = 0 THEN B%(0) = X - 1:B%(B%(0) + 1) = M + 1:A(0) = M: GOTO 410
186 N = 0
190 M = M + 1:N = N + 1
195  IF M = 1000 THEN  PRINT "CAN'T ACCEPT ANY MORE DATA":I(X) =  - 1: GOTO 410
200  PRINT "MEASUREMENT ";N;: INPUT A(M),P1: IF A(M) = 0 THEN  PRINT : GOTO 250
210  IF  ABS (A(M) - P1) < 5 THEN 220
212  PRINT : PRINT "** WARNING:  THE DIFFERENCE BETWEEN THE": PRINT "LAST TWO ENTRIES IS GREA
     TER THAN 5": PRINT : PRINT "IS THAT O.K. (Y,N)";: INPUT A$: PRINT : IF A$ = "Y" THEN 220

214  GOTO 200
220  IF A(M) = 999 THEN  PRINT : PRINT "** SAMPLE DELETED **": PRINT :M = O8: GOTO 180
230 A(M) = (A(M) + P1) / 2: IF A(M) > A(M - 1) THEN 190
238  PRINT : PRINT "** ORDER OF DATA JUST ENTERED IS IN-": PRINT "CORRECT.  RE-ENTER THIS SAM
     PLE":M = O8: PRINT : GOTO 180
250 T = A(M - 1):P = M - 1:S(X) = A(M - 1):M = M - 2
300  IF M = 0 THEN 350
310  IF A(M) = 0 THEN 350
320 A(M) = A(M) / T * 100:M = M - 1: GOTO 300
350 A(P) = 0:M = P:X = X + 1:B%(X) = M + 1: GOTO 180
```

```
370  REM *
371  REM *     AFTER DATA HAS BEEN ADDED TO THE FILE, ALLOW OPTION OF ORDERING THE DATA    *
372  REM *
410  GOSUB 9999: PRINT : PRINT "DO YOU WANT THE DATA PUT IN ORDER": PRINT "(Y,N)";: INPUT A$:
      IF A$ = "N" THEN 10010
411  GOSUB 1025: GOTO 10010
999  REM *
1000  REM *     ROUTINE TO ALLOW ORDERING OF SAMPLES BY DECREASING NUMBER OF SITES          *
1001  REM *
1025 M3 = B%(0):D(C%(M3 + 1)) =  - 1:C%(M3 + 1) = B%(M3 + 1): FOR X = 1 TO M3:B(X) = B%(X + 1
     ) - B%(X): NEXT X
1027 Q = 0:L = 0: FOR X = 10 TO 1 STEP  - 1: FOR X1 = 1 TO M3: IF B(X1) <  > X THEN 1100
1050 L = L + 1:C%(L) = Q + 1:I1(L) = I(X1):S1(L) = S(X1): FOR Z = B%(X1) TO B%(X1 + 1) - 1:Q =
     Q + 1:D(Q) = A(Z): NEXT Z:B(X1) = 0
1100  NEXT X1: NEXT X: FOR X = 1 TO M3:B%(X) = C%(X):I(X) = I1(X):S(X) = S1(X): NEXT X
1116  FOR X = 1 TO Q:A(X) = D(X): NEXT X:A(0) = Q:B%(0) = M3:B%(B%(0) + 1) = Q + 1: RETURN
1118  REM *
1119  REM *     ROUTINE TO PRINT OUT ALL SAMPLES IN THE DATA SET                             *
1120  REM *
1123  GOSUB 9999: PRINT : PRINT "HARDCOPY (Y,N)";: INPUT A$
1125  IF A$ = "Y" THEN  PR# 1: PRINT  CHR$ (28);
1126  PRINT : GOSUB 9999: PRINT "FILENAME: ";F$
1127  IF A$ = "N" THEN  SPEED= 180
1130  FOR M = 1 TO 200:B(M) = 0: NEXT M: FOR M = 1 TO 10:L(M) = 0: NEXT M:M3 = B%(0): PRINT :
      GOSUB 9999: PRINT : FOR M = 1 TO M3
1132 BM = B%(M + 1) - B%(M):L(BM - 1) = L(BM - 1) + 1: NEXT M: PRINT "TOTAL NUMBER OF MOLECUL
     ES = ";M3: PRINT
1134  FOR M = 1 TO 25: IF L(M) <  > 0 THEN  PRINT "MOLECULES WITH ";M;" BUSHES = ";L(M)
1136  NEXT M: PRINT : GOSUB 9999
1200  FOR X = 1 TO M3: PRINT "SAMPLE # ";I(X),"SIZE= ";S(X)
1300  FOR Y = B%(X) TO B%(X + 1) - 2: PRINT  INT ((A(Y) + .05) * 10) / 10;"  ";: NEXT Y: PRINT
     : GOSUB 9999: NEXT X: PRINT : IF A$ = "Y" THEN  PR# 0
1520  SPEED= 255: RETURN
1900  REM *
1901  REM *     ROUTINE TO SAVE DATA TO A DISK FILE                                          *
1902  REM *
2000  PRINT : PRINT "WOULD YOU LIKE TO SAVE THIS DATA": PRINT "(Y,N)";: INPUT A$: IF A$ = "N"
      THEN  RETURN
2001  PRINT : PRINT "FILENAME";: INPUT A$: PRINT
2004 D$ = "": PRINT D$;"OPEN";A$: PRINT D$;"WRITE";A$
2010  FOR X = 1 TO 40: PRINT E(X): NEXT X: PRINT A(0): FOR X = 1 TO A(0): PRINT A(X): NEXT X:
      PRINT B%(0): FOR X = 1 TO B%(0): PRINT B%(X): NEXT X: FOR X = 1 TO B%(0): PRINT I(X): NEXT
     X: FOR X = 1 TO B%(0): PRINT S(X): NEXT X
2024  PRINT D$;"CLOSE";A$: PRINT D$;"LOCK";A$: PRINT D$
2030  PRINT : PRINT "DATA HAS BEEN SAVED": PRINT : PRINT "FILENAME= ";A$: PRINT : PRINT : RETURN

2900  REM *
2901  REM *     ROUTINE TO READ DATA FROM A DISK FILE                                        *
2902  REM *
3000  PRINT "FILENAME";: INPUT A$:F$ = A$
```

```
3004 D$ = "": PRINT D$;"OPEN";A$: PRINT D$;"READ";A$
3010  FOR X = 1 TO 40: INPUT E(X): NEXT X: INPUT A(0): FOR X = 1 TO A(0): INPUT A(X): NEXT X:
      INPUT B%(0):M = B%(0):M3 = B%(0): FOR X = 1 TO B%(0): INPUT B%(X): NEXT X: FOR X = 1 TO
     B%(0): INPUT I(X): NEXT X: FOR X = 1 TO B%(0): INPUT S(X): NEXT X
3011  PRINT D$:B%(B%(0) + 1) = A(0) + 1
3020  GOSUB 9999: PRINT "DATA DESCRIPTION:": PRINT
3030  FOR X = 1 TO 40: PRINT  CHR$ (E(X));: NEXT X: PRINT
3040  RETURN
5900  REM *
5901  REM *     ROUTINE TO HANDLE DISK I/O ERRORS                                      *
5902  REM *
6000  IF  PEEK (222) = 255 THEN  END
6001  IF  PEEK (222) = 9 THEN  PRINT "DISK FULL - TRY ANOTHER DISK": PRINT "DELETE";A$: PRINT
     : GOSUB 2001: END
6005  IF  PEEK (222) < > 5 THEN 6050
6007  PRINT : PRINT "";A$;" NOT FOUND ON DISK": PRINT "CLOSE";A$: PRINT "DELETE";A$
6035  PRINT : PRINT "CATALOG LISTING (Y,N)";: INPUT A$: IF A$ = "N" THEN  PRINT : GOTO 140
6037  IF A$ < > "Y" THEN 6035
6039  PRINT "CATALOG": PRINT : GOTO 139
6050  IF  PEEK (222) < > 10 AND  PEEK (222) < > 254 THEN  POKE 216,0: RESUME
6060  PRINT "";A$;" ALREADY IN CATALOG": PRINT : PRINT "OVERWRITE (Y,N)";: INPUT C$: IF C$ =
     "N" THEN 10045
6090  PRINT "UNLOCK";A$: PRINT "DELETE";A$: GOSUB 2004: END
7000  REM *
7001  REM *     ROUTINE TO PRINT A DASHED LINE                                         *
7002  REM *
9999  FOR X0 = 1 TO 39: PRINT "-";: NEXT X0: PRINT : RETURN
10007  REM *
10008  REM *     BRANCHING POINT FOR MAJOR OPTIONS OF THIS PROGRAM                     *
10009  REM *
10010  GOSUB 9999: PRINT : PRINT "TO MODIFY THIS FILE YOU MAY:": PRINT
10011  PRINT "     1) ADD SAMPLES TO THE FILE"
10012  PRINT
10013  PRINT "     2) DELETE SAMPLES FROM THE FILE"
10020  PRINT
10030  PRINT "     3) PRINT OUT THE DATA"
10032  PRINT
10034  PRINT "     4) STORE THE DATA ON THE DISK": PRINT "         AND STOP"
10035  PRINT
10037  GOSUB 9999
10040  PRINT "ADD, DELETE, PRINT OR STOP": PRINT "(A,D,P,S)";: INPUT B$
10045  IF B$ = "S" THEN  GOSUB 9999: PRINT : GOSUB 2000: END
10050  IF B$ = "A" THEN X = B%(0) + 1:M = A(0):M = A(0):B%(X) = M + 1:M3 = B%(0): HOME : GOTO
     165
10051  IF B$ = "P" THEN  GOSUB 1123: GOTO 10010
10052  IF B$ < > "D" THEN 10035
10053  REM *
10054  REM *     ROUTINE TO DELETE SAMPLES FROM THE DATA SET                           *
10055  REM *
10059  PRINT : GOSUB 9999: PRINT "ENTER SAMPLES TO BE DELETED BELOW:": PRINT : PRINT
```

```
10060  PRINT "SAMPLE #";: INPUT I: PRINT
10066  IF I = 0 THEN 10010
10070  FOR X = 1 TO B%(0): IF I(X) = I THEN 10090
10075  NEXT X: PRINT "SAMPLE NOT FOUND": PRINT : GOTO 10060
10090 L = B%(X + 1) - B%(X): IF X = B%(0) THEN 10130
10100  FOR Y = X TO B%(0) - 1:I(Y) = I(Y + 1):S(Y) = S(Y + 1): NEXT Y: FOR Y = B%(X) TO A(0) -
     L:A(Y) = A(Y + L): NEXT Y
10120  FOR Y = X TO B%(0):B%(Y) = B%(Y) - L: NEXT Y: FOR Y = X TO B%(0) - 1:B%(Y) = B%(Y + 1)
     : NEXT Y
10130 B%(0) = B%(0) - 1:A(0) = A(0) - L:B%(B%(0) + 1) = A(0) + 1: GOTO 10060
```

```
1  REM ************************************************************************
2  REM *     DNA2  -  DNA ANALYSIS PROGRAM     -LAST REVISED: 12/19/78           *
3  REM *     SCOTT MC COURT                    -DEPT. OF COMPUTER SCIENCE        *
4  REM ************************************************************************
5  LOMEM: 24576: PRINT "NOMONI,0,C": HOME : SPEED= 255: ONERR  GOTO 6000
10  GOSUB 9999: PRINT  TAB( 14);"DNA ANALYSIS": GOSUB 9999
110  DIM A(1000),B%(201),I(201),S(201),D(15),E(40),H(102),T(102)
114  GOSUB 2000
116  PRINT : PRINT "ARRANGEMENT, HISTOGRAM OR CORRELATIONS": PRINT "(A,H,C)";: INPUT A$: IF A
    $ = "H" THEN 20020
117  IF A$ = "C" THEN 13000
118  IF A$ <  > "A" THEN 116
119  REM *
120  REM *     ROUTINE TO ORIENT SAMPLES BASED ON USER PARAMETERS                   *
121  REM *
163 AN = 0: FOR X = 1 TO B%(0):I(X) =  ABS (I(X)): NEXT X
165  PRINT : PRINT "DO YOU WANT TO PUT CERTAIN SAMPLES": PRINT "   FIRST (Y,N)";: INPUT A$: IF
    A$ = "N" THEN 179
177  IF A$ <  > "Y" THEN 165
178  GOSUB 5000
179  PRINT : PRINT "INTERVAL SIZE (1-100)";: INPUT G8: IF G8 < 1 OR G8 > 100 THEN 179
180 G9 =  INT (100 / G8): FOR X = 1 TO G9:H(X) = 0: NEXT X
182  PRINT : PRINT "DO YOU WANT TO DO A BEST HISTOGRAM FIT": PRINT "   ON THE FIRST FEW SAMPL
    ES (Y,N)";: INPUT A$: IF A$ = "N" THEN N4 = 1:E = 1:K4 = 1:M9 = 0: GOSUB 1500: GOTO 186
184  IF A$ <  > "Y" THEN 182
185  GOSUB 7000:K4 = E
186  PRINT : PRINT "MINIMUM AND MAXIMUM BUSH NUMBER";: INPUT V1,V9: IF V9 < V1 THEN 186
187  PRINT : PRINT "CORRELATION CUTOFF VALUE";: INPUT E1: IF E = 1 THEN  PRINT : PRINT "START
     ANALYSIS IN THE MIDDLE (Y,N)";: INPUT A$: IF A$ = "Y" THEN  PRINT : PRINT "START WITH W
    HICH SAMPLE";: INPUT E:E = E - 1: FOR X = 1 TO G9:H(X) = 0: NEXT X: FOR N4 = 1 TO E: GOSUB
    1500: NEXT N4
188 K4 = E
195  PRINT : PRINT "THE ANALYSIS INCLUDES THE FOLLOWING": PRINT "   SAMPLES:": PRINT : FOR X =
    1 TO E: PRINT I(X),: NEXT X
534  FOR K3 = E + 1 TO B%(0):G6 =  PEEK ( - 16384): IF G6 = 211 THEN  POKE  - 16368,0: PRINT
    : PRINT "  ** ANALYSIS TERMINATED **": GOTO 116
537  IF B%(K3 + 1) - B%(K3) - 1 > V9 OR B%(K3 + 1) - B%(K3) - 1 < V1 THEN I(K3) =  -  ABS (I(
    K3)): GOTO 640
538  IF AN = 1 AND I(K3) >  = 0 THEN 560
539 NB = B%(K3 + 1) - B%(K3) - 1:N4 = K3: GOSUB 1200:F1 = C: GOSUB 660: GOSUB 1200:R1 = C
547  IF F1 / K4 / NB <  = E1 AND R1 / K4 / NB <  = E1 THEN RE = 1: PRINT  ABS (I(K3));"*",:I(
    K3) =  - ( ABS (I(K3))): GOTO 640
549  IF R1 < F1 THEN  GOSUB 660
558  GOSUB 1500:K4 = K4 + 1:I(K3) =  ABS (I(K3))
560  PRINT I(K3),
640  NEXT K3: PRINT : GOSUB 1000: PRINT : PRINT "HISTOGRAM COEFFICIENT= "; INT ((C + .005) *
    100) / 100: IF RE = 0 THEN 20000
642  PRINT : PRINT "SOME SAMPLES WERE LEFT OUT OF THIS": PRINT "ARRANGEMENT.  RE-ANALYZE (Y,N
    )";: INPUT A$: IF A$ = "N" THEN 20000
```

```
643 AN = 1:RE = 0: PRINT : PRINT "CHANGE CORRELATION CUTOFF VALUE (Y/N)";: INPUT A$: IF A$ =
    "Y" THEN  PRINT : PRINT "NEW CUTOFF VALUE";: INPUT E1
644  PRINT : PRINT "REANALYSIS:": GOTO 195
645  REM *
646  REM *     SUBROUTINE TO REVERSE ORIENTATION OF SAMPLE NUMBER 'N4'                         *
647  REM *
660 XA = B%(N4):XB = B%(N4 + 1) - 2: FOR X9 = 0 TO  INT ((XB - XA) / 2):T = A(XA + X9):A(XA +
    X9) = 100 - A(XB - X9):A(XB - X9) = 100 - T: NEXT X9: RETURN
665  REM *
666  REM *     SUBROUTINE TO CALCULATE THE HISTOGRAM COEFFICIENT BASED ON THE CURRENT          *
667  REM *     HISTOGRAM                                                                       *
668  REM *
1000 C = 0:C8 = 0: FOR P3 = 1 TO G9:C = C + H(P3) ↑ 2:C8 = C8 + H(P3): NEXT P3:C = C / C8: RETURN

1005  REM *
1006  REM *     SUBROUTINE TO DETERMINE HOW WELL MOLECULE NUMBER 'N4', IN ITS CURRENT          *
1007  REM *     ORIENTATION, FITS INTO THE PRESENT HISTOGRAM                                   *
1008  REM *
1200 C = 0: FOR B9 = B%(N4) TO B%(N4 + 1) - 2:B0 =  INT (A(B9) / G8) + 1:C = C + H(B0): IF B0
     > 1 THEN C = C + .3413 * (H(B0 + 1) + H(B0 - 1))
1203  IF B0 > 2 THEN C = C + .1359 * (H(B0 + 2) + H(B0 - 2))
1204  NEXT B9:C =  INT ((C + .005) * 100) / 100: RETURN
1250  REM *
1251  REM *     SUBROUTINE TO REMOVE MOLECULE 'N4' (SMOOTHED) FROM THE PRESENT HISTOGRAM    *
1252  REM *
1300  FOR B9 = B%(N4) TO B%(N4 + 1) - 2:B0 =  INT (A(B9) / G8) + 1:H(B0) = H(B0) - 1:H(B0 + 1
     ) = H(B0 + 1) - .3413:H(B0 + 2) = H(B0 + 2) - .1359: IF B0 > 1 THEN H(B0 - 1) = H(B0 - 1
     ) - .3413
1310  IF B0 > 2 THEN H(B0 - 2) = H(B0 - 2) - .1359
1320  NEXT B9: RETURN
1350  REM *
1351  REM *     SUBROUTINE TO ADD MOLECULE 'N4' (SMOOTHED) INTO THE PRESENT HISTOGRAM         *
1352  REM *
1400  FOR B9 = B%(N4) TO B%(N4 + 1) - 2:B0 =  INT (A(B9) / G8) + 1:H(B0) = H(B0) + 1:H(B0 + 1
     ) = H(B0 + 1) + .3413:H(B0 + 2) = H(B0 + 2) + .1359: IF B0 > 1 THEN H(B0 - 1) = H(B0 - 1
     ) + .3413
1410  IF B0 > 2 THEN H(B0 - 2) = H(B0 - 2) + .1359
1420  NEXT B9: RETURN
1450  REM *
1451  REM *     SUBROUTINE TO ADD MOLECULE 'N4' (UNSMOOTHED) INTO THE PRESENT HISTOGRAM     *
1452  REM *
1500  FOR B9 = B%(N4) TO B%(N4 + 1) - 2:B0 =  INT (A(B9) / G8) + 1:H(B0) = H(B0) + 1: NEXT B9
     : RETURN
1550  REM *
1551  REM *     SUBROUTINE TO REMOVE MOLECULE 'N4' (UNSMOOTHED) FROM THE PRESENT HISTOGRAM *
1552  REM *
1600  FOR B9 = B%(N4) TO B%(N4 + 1) - 2:B0 =  INT (A(B9) / G8) + 1:H(B0) = H(B0) - 1: NEXT B9
     : RETURN
```

```
1650  REM *
1651  REM *     SUBROUTINE TO READ IN A DATA FILE FROM THE DISK                            *
1652  REM *
2000  PRINT : PRINT "FILENAME";: INPUT A$:C$ = A$:D$ = ""
2005  PRINT D$;"CLOSE": PRINT D$;"OPEN";A$: PRINT D$;"READ";A$
2010  FOR X = 1 TO 40: INPUT E(X): NEXT X: INPUT A(0): FOR X = 1 TO A(0): INPUT A(X): NEXT X:
      INPUT B%(0):B%(B%(0) + 1) = A(0) + 1: FOR X = 1 TO B%(0): INPUT B%(X): NEXT X: FOR X =
     1 TO B%(0): INPUT I(X): NEXT X: FOR X = 1 TO B%(0): INPUT S(X): NEXT X
2024  PRINT D$: PRINT : GOSUB 9999: PRINT : PRINT "DATA DESCRIPTION:": PRINT : FOR X = 1 TO 4
     0: PRINT  CHR$ (E(X));: NEXT X: PRINT : PRINT : PRINT "NO. MOLECULES= ";B%(0): PRINT : GOSUB
     9999: RETURN
2100  REM *
2101  REM *     SUBROUTINE TO SMOOTH THE PRESENT HISTOGRAM                                 *
2102  REM *
3360  FOR X = 1 TO G9:T(X) = 0: NEXT X: FOR X = 2 TO G9 - 1:T(X - 1) = T(X - 1) + .3413 * H(X
     ):T(X + 1) = T(X + 1) + .3413 * H(X):T(X - 2) = T(X - 2) + .1359 * H(X):T(X + 2) = T(X +
     2) + .1359 * H(X): NEXT X: FOR X = 1 TO G9:H(X) = H(X) + T(X): NEXT X: RETURN
3370  REM *
3371  REM *     SUBROUTINE TO ALLOW SPECIFIC SAMPLES TO BE MOVED TO THE TOP OF THE DATA SET*
3372  REM *
5000  PRINT : PRINT "HOW MANY MOLECULES (1-10)";: INPUT N: PRINT : PRINT "INPUT SAMPLE NUMBER
     S BELOW:": FOR X = 1 TO N
5005  PRINT : PRINT "MOLECULE #";X;: INPUT X1: FOR Y = 1 TO B%(0): IF I(Y) = X1 THEN 5100
5080  NEXT Y: PRINT : PRINT "SAMPLE # ";X1;" NOT FOUND": GOTO 5005
5100  IF Y < X THEN  PRINT : PRINT "SAMPLE # ";X1;" HAS ALREADY BEEN MOVED": GOTO 5005
5102  IF Y = X THEN 5250
5105 L = B%(Y + 1) - B%(Y): FOR Z = B%(Y) TO B%(Y) + L - 1:D(Z - B%(Y) + 1) = A(Z): NEXT Z: FOR
     Z = B%(Y + 1) - 1 TO B%(X) + L STEP  - 1:A(Z) = A(Z - L): NEXT Z: FOR Z = B%(X) TO B%(X)
      + L - 1:A(Z) = D(Z - B%(X) + 1): NEXT Z
5110 B = X: FOR Z = B%(X) TO B%(Y): IF A(Z) = 0 THEN B = B + 1:B%(B) = Z + 1
5115  NEXT Z
5117 T1 = I(Y):T2 = S(Y): FOR Z = Y TO X + 1 STEP  - 1:I(Z) = I(Z - 1):S(Z) = S(Z - 1): NEXT
     Z:I(X) = T1:S(X) = T2
5250  NEXT X: RETURN
5260  REM *
5261  REM *     ROUTINE TO HANDLE DISK I/O ERRORS                                          *
5262  REM *
6000  IF  PEEK (222) = 255 THEN  END
6001  IF  PEEK (222) = 9 THEN  PRINT "DISK FULL - TRY ANOTHER DISK": PRINT "DELETE";A$: PRINT
     : GOTO 10025
6005  IF  PEEK (222) < > 5 THEN 6050
6007  PRINT : PRINT "";A$;" NOT FOUND ON DISK": PRINT : PRINT "CLOSE";A$: PRINT "DELETE";A$
6035  PRINT "CATALOG LISTING (Y,N)";: INPUT A$: IF A$ = "N" THEN  PRINT : GOTO 114
6037  IF A$ < > "Y" THEN 6035
6039  PRINT "CATALOG": PRINT : PRINT : GOTO 114
6050  IF  PEEK (222) < > 10 THEN  POKE 216,0: RESUME
6060  PRINT "";A$;" ALREADY IN CATALOG": PRINT : PRINT "OVERWRITE (Y,N)";: INPUT B$: PRINT : IF
     B$ = "N" THEN 10025
6090  PRINT "UNLOCK";A$: PRINT "DELETE";A$: GOTO 10028
```

```
6120  REM *
6121  REM *     SUBROUTINE TO FORM BEST POSSIBLE HISTOGRAM CONSISTING OF THE FIRST 'E'      *
6122  REM *     SAMPLES                                                                     *
6123  REM *
7000  PRINT : PRINT "HOW MANY SAMPLES (1-10)";: INPUT E:M9 = 0:M8 = 0: FOR N4 = 1 TO E: GOSUB
    1400: NEXT N4: GOSUB 1000:M9 = C:IC = E - 1
7010  FOR X3 = 1 TO 2 ↑ IC: FOR Y = 0 TO IC - 1: IF X3 / 2 ↑ Y =  INT (X3 / 2 ↑ Y) THEN N4 =
    Y + 1: GOSUB 1300: GOSUB 660: GOSUB 1400
7070  NEXT Y: GOSUB 1000: IF C > M9 THEN M9 = C:M8 = X3
7075  NEXT X3:XY = M8: FOR X = IC - 1 TO 0 STEP  - 1:J = XY - 2 ↑ X: IF J >  = 0 THEN N4 = X +
    1: GOSUB 660:XY = J
7080  NEXT X
7085  FOR X = 1 TO G9:H(X) = 0: NEXT X: FOR N4 = 1 TO E: GOSUB 1500: NEXT N4
7139  PRINT : PRINT "WOULD YOU LIKE THE BEST-FIT ARRANGEMENT": PRINT "  PRINTED OUT (Y,N)";: INPUT
    A$: IF A$ = "N" THEN  RETURN
7155  PR# 1: FOR X = 1 TO E: GOSUB 9999: PRINT "SAMPLE #";I(X): FOR Y = B%(X) TO B%(X + 1) -
    2: IF  INT (A(Y) / 10) = 0 THEN  PRINT  TAB( 1); INT (A(Y));" ";: GOTO 7190
7180  PRINT  TAB(  INT (A(Y) / 10) * 4); INT (A(Y));" ";
7190  NEXT Y: PRINT : NEXT X: GOSUB 9999: PRINT : PR# 0: RETURN
7200  REM *
7201  REM *     SUBROUTINE TO PRINT A DASHED LINE
7203  REM *
9999  FOR HW = 1 TO 39: PRINT "-";: NEXT HW: PRINT : RETURN
10000  REM *
10001  REM *     SUBROUTINE TO SAVE THE DATA IN ITS CURRENT ARRANGEMENT TO THE DISK
10002  REM *
10017  PRINT : PRINT "WOULD YOU LIKE TO SAVE THE CURRENT": PRINT "  DATA (Y,N)";: INPUT A$: IF
    A$ = "N" THEN  RETURN
10018  IF A$ <  > "Y" THEN 10017
10019  GOSUB 9999: FOR X = 1 TO 40:E(X) = 0: NEXT X
10021  PRINT : PRINT "INPUT DATA DESCRIPTION (ONE LINE)": PRINT : INPUT A$: IF  LEN (A$) > 40
     THEN  PRINT : PRINT "TOO LONG.  RE-ENTER.": GOTO 10021
10023  FOR X = 1 TO  LEN (A$):E(X) =  ASC ( MID$ (A$,X,1)): NEXT X:E(0) =  LEN (A$)
10025  PRINT : PRINT "FILENAME";: INPUT A$:D$ = "": PRINT
10028  PRINT D$;"OPEN";A$
10030  PRINT D$;"WRITE";A$
10040  FOR X = 1 TO 40: PRINT E(X): NEXT X: PRINT A(0): FOR X = 1 TO A(0): PRINT A(X): NEXT X
    : PRINT B%(0): FOR X = 1 TO B%(0): PRINT B%(X): NEXT X: FOR X = 1 TO B%(0): PRINT I(X): NEXT
    X: FOR X = 1 TO B%(0): PRINT S(X): NEXT X
10050  PRINT D$;"CLOSE";A$
10052  PRINT D$;"LOCK";A$
10054  PRINT D$
10060  PRINT "DATA HAS BEEN SAVED"
10070  REM *
10071  REM *     BRANCHING POINT FOR MAJOR OPTIONS AFTER ANALYZING DATA
10072  REM *
11000  PRINT : GOSUB 9999: PRINT "YOU MAY NOW:": PRINT : PRINT "     1) CONTINUE WITH CURRENT
     DATA": PRINT : PRINT "     2) READ IN SOME OTHER DATA": PRINT : PRINT "     3) SAVE CUR
    RENT ARRANGEMENT": PRINT : PRINT "     4) TERMINATE PROCESSING": GOSUB 9999
11080  PRINT "FUNCTION";: INPUT A: IF A < 1 OR A > 4 THEN 11080
```

```
11090  IF A = 1 THEN 116
11100  IF A = 2 THEN  GOSUB 10000: GOTO 114
11110  IF A = 3 THEN  GOSUB 10019: GOTO 11000
11120  IF A = 4 THEN  GOSUB 10000: END
11130  REM *
11131  REM *     ROUTINE TO CALCULATE SIMILARITIES BETWEEN ALL POSSIBLE PAIRS OF MOLECULES *
11132  REM *
13000  GOSUB 9999: PRINT "INTERVAL SIZE (1-100)";: INPUT G8:G9 =  INT (100 / G8): FOR X = 1 TO
     G9:H(X) = 0: NEXT X
13010  PRINT : PRINT "MINIMUM AND MAXIMUM BUSH NUMBER";: INPUT M1,M9: IF M9 < M1 THEN 13010
13020  PRINT : PR# 1: PRINT  CHR$ (28);: PRINT : GOSUB 9999: PRINT "MOL/BUSHES"; TAB( 14);"MO
     L/BUSHES"; TAB( 27);"CORRELATION": GOSUB 9999: PRINT  TAB( 27);"F"; TAB( 32);"R"; TAB( 3
     7);"D": PRINT  TAB( 26);"-------------"
13490  FOR X = 1 TO 100:H(X) = 0: NEXT X: FOR Y = 1 TO BX(0): IF BX(Y + 1) - BX(Y) - 1 < M1 OR
     BX(Y + 1) - BX(Y) - 1 > M9 THEN 13895
13530 N4 = Y: GOSUB 1500: FOR X = Y + 1 TO BX(0): IF BX(X + 1) - BX(X) - 1 < M1 OR BX(X + 1) -
     BX(X) - 1 > M9 THEN 13890
13630 G6 =  PEEK ( - 16384): IF G6 = 211 THEN  POKE  - 16368,0: PRINT : PRINT "** ANALYSIS TE
     RMINATED **": PRINT : PR# 0: GOTO 13897
13635 N4 = X: GOSUB 1200:F1 = C:N4 = X: GOSUB 660: GOSUB 1200:R1 = C:C =  ABS (F1 - R1): IF C
      > 1 THEN  PRINT  TAB( 4);I(Y);"/";BX(Y + 1) - BX(Y) - 1; TAB( 17);I(X);"/";BX(X + 1) -
     BX(X) - 1; TAB( 27);F1; TAB( 32);R1; TAB( 37);C
13637  IF C > 1 THEN  PRINT  TAB( 26); INT ((F1 / (BX(X + 1) - BX(X) - 1) * 100) + .005) / 10
     0; TAB( 31); INT ((R1 / (BX(X + 1) - BX(X) - 1) * 100) + .005) / 100; TAB( 36); INT ((C /
     (BX(X + 1) - BX(X) - 1) * 100) + .005) / 100
13639  IF C > 1 THEN  PRINT  TAB( 26);"-------------"
13890  NEXT X:N4 = Y: GOSUB 1600
13895  NEXT Y
13897  PRINT : GOSUB 9999: PR# 0
13900  PRINT : PRINT "WOULD YOU LIKE CERTAIN SAMPLES PRINTED": PRINT "    ON THE SCREEN (Y,N)"
     ;: INPUT A$: IF A$ = "N" THEN 11000
13918  GOSUB 9999: PRINT : PRINT "INPUT SAMPLE NUMBERS: (TYPE '0' TO STOP)": FOR N = 1 TO 25:
      INPUT T(N): IF T(N) = 0 THEN N = N - 1: GOTO 13944
13930  NEXT N
13944  IF N = 0 THEN 13900
13945  PR# 1: GOSUB 9999: FOR X = 1 TO N:I = T(X): FOR N1 = 1 TO BX(0): IF I(N1) <  > I THEN
     13970
13947  PRINT I
13950  FOR M = BX(N1) TO BX(N1 + 1) - 2: PRINT  INT (A(M));" ";: NEXT M: PRINT "  /  ";:N4 =
     N1: GOSUB 660: FOR M = BX(N1) TO BX(N1 + 1) - 2: PRINT  INT (A(M));" ";: NEXT M: PRINT :
      GOSUB 9999: GOTO 13985
13970  NEXT N1: PRINT : PRINT "SAMPLE #";I;" NOT FOUND": GOSUB 9999
13985  NEXT X: PR# 0: GOTO 13900
14000  REM *
14001  REM *     ROUTINE TO GENERATE A HISTOGRAM ON THE VIDEO MONITOR BASED ON THE           *
14002  REM *     CURRENT DATA ARRANGEMENT                                                    *
14003  REM *
20000  PRINT : PRINT "WOULD YOU LIKE A HISTOGRAM (Y,N)";: INPUT A$: IF A$ = "N" THEN 11000
20010  IF A$ <  > "Y" THEN 20000
20020  PRINT : GOSUB 9999
```

```
20021  PRINT : PRINT "HISTOGRAM STEP SIZE (1-100)";: INPUT G8: IF G8 < 0 OR G8 > 100 THEN 200
    21
20025  IF G8 = 0 THEN 11000
20030 G9 =  INT (100 / G8): FOR X = 1 TO G9:H(X) = 0: NEXT X
20060  PRINT : PRINT "MULTIPLICATION FACTOR";: INPUT R4: IF R4 > 1 THEN 20060
20070  PRINT : PRINT "MINIMUM AND MAXIMUM BUSH NUMBER";: INPUT C1,C9: IF C1 > C9 THEN 20070
20080 I9 = 5: PRINT : PRINT "REVERSE (Y,N)";: INPUT A$: IF A$ = "Y" THEN  FOR N4 = 1 TO B%(0)
     : GOSUB 660: NEXT N4: GOTO 20100
20090  IF A$ < > "N" THEN 20080
20100  FOR Y = 1 TO B%(0): IF B%(Y + 1) - B%(Y) - 1 < C1 OR B%(Y + 1) - B%(Y) - 1 > C9 THEN 2
     0200
20105  IF I(Y) < 0 THEN 20200
20120  FOR X = B%(Y) TO B%(Y + 1) - 2:B0 =  INT ((A(X) * R4) / G8) + 1: IF B0 > G9 THEN B0 =
     G9
20150 H(B0) = H(B0) + 1: NEXT X
20200  NEXT Y
20210  PRINT : PRINT "SMOOTHING (Y,N)";: INPUT B$: IF B$ = "Y" THEN  GOSUB 3360: GOTO 20260
20220  IF B$ < > "N" THEN 20210
20260 M9 = 0: FOR X = 1 TO G9: IF H(X) > M9 THEN M9 = H(X)
20280  NEXT X
20290  IF M9 = 0 THEN M9 = 1
20300 L3 =  INT (189 / M9):L2 =  INT (360 / M9):I8 =  - 1:M = 0: HGR2 : HCOLOR= 3:L =  INT (2
     70 / G9):B =  INT ((270 - (G9 * L)) / 2):N = 0
20320  FOR X = B + 1 TO B + (G9 * L) STEP L:N = N + 1:I8 = I8 + 1: FOR Y = 0 TO L - 1: IF I8 =
     I9 THEN I8 = 0: HPLOT X + Y,191 TO X + Y,189
20330  IF H(N) < > 0 THEN  HPLOT X + Y,189 TO X + Y,189 - H(N) * L3
20340  NEXT Y: NEXT X: FOR X = 0 TO 189 STEP L3: HPLOT B + 1,189 - X TO B + (G9 * L),189 - X:
      NEXT X
20350  GET A$
20352  IF A$ < > "A" OR RE < > 1 THEN 20360
20355  IF A$ = "A" AND RE = 1 THEN  TEXT : GOSUB 9999: PRINT : PRINT "ANALYSIS CONTINUES:": FOR
     X = 1 TO G9:H(X) = 0: NEXT X: FOR X = 1 TO B%(0): IF I(X) > 0 THEN N4 = X: GOSUB 1500
20356  NEXT X: GOTO 642
20360  IF A$ =  CHR$ (13) THEN  TEXT : GOTO 20390
20370  IF A$ = "R" THEN  FOR X = 1 TO  INT (G9 / 2):T = H(X):H(X) = H(G9 + 1 - X):H(G9 + 1 -
     X) = T: NEXT X: GOTO 20300
20380  IF A$ = "C" THEN  GOSUB 30000: GOTO 20350
20385  GOTO 20350
20386  REM *
20387  REM *     ROUTINE TO GENERATE 3 DIFFERENT TYPES OF STATISTICS BASED ON THE CURRENT  *
20388  REM *     HISTOGRAM                                                                 *
20389  REM *
20390  PRINT : PRINT "WOULD YOU LIKE STATISTICS (Y,N)";: INPUT A$: IF A$ = "N" THEN 20020
20400  IF A$ < > "Y" THEN 20390
20405  GOSUB 9999: PRINT "  1) PRINT STATISTICS ON BUSHES WITHIN": PRINT "      A SINGLE RANGE
     ": PRINT "  2) PRINT ALL MOLECULES WITH A BUSH": PRINT "      IN EACH OF A NUMBER OF RANG
     ES"
20406  PRINT "  3) PRINT A DISTRIBUTION OF THE NUMBER": PRINT "      OF BUSHES IN A GIVEN RANG
     E FOR ALL": PRINT "      MOLECULES": GOSUB 9999
20407  PRINT : PRINT "WHICH TYPE (1-3)";: INPUT T: IF T < 1 OR T > 3 THEN 20407
```

```
20408  PR# 1: GOSUB 30400: PR# 0: GOSUB 9999: ON T GOTO 20415,20720,20900
20409  REM *
20410  REM *     STATISTICS:  TYPE 1
20411  REM *
20415  PRINT : PRINT "INPUT RANGE (MIN,MAX)";: INPUT R1,R2: PRINT : IF R1 = 0 AND R2 = 0 THEN
     20390
20420 SU = 0:NO = 0:SX = 0: FOR X = 1 TO B%(0): IF B%(X + 1) - B%(X) - 1 < C1 OR B%(X + 1) -
     B%(X) - 1 > C9 THEN 20600
20422  IF I(X) < 0 THEN 20600
20425  FOR Y = B%(X) TO B%(X + 1) - 2: IF A(Y) * R4 < R1 OR A(Y) * R4 > R2 THEN 20590
20440 SU = A(Y) * R4 + SU:SX = SX + (A(Y) * R4) ↑ 2:NO = NO + 1
20590  NEXT Y
20600  NEXT X
20606  PR# 1: GOSUB 9999: PRINT : PRINT "RANGE= ";R1;" - ";R2: PRINT : PRINT "NUMBER OF MOLEC
     ULES= ";NO: IF NO = 0 THEN  PR# 0: PRINT : GOSUB 9999: GOTO 20415
20630  PRINT : PRINT "MEAN MEASUREMENT= ";SU / NO:SD = (SX - ((SU ↑ 2) / NO)) / NO: PRINT : PRINT
     "STD. DEV.= "; SQR (SD): PRINT : PRINT "VARIANCE= ";SD: PRINT : PR# 0: GOSUB 9999: GOTO
     20415
20717  REM *
20718  REM *     STATISTICS:  TYPE 2
20719  REM *
20720  PRINT : PRINT "NUMBER OF SITES";: INPUT NO: IF NO = 0 THEN 20390
20730  FOR X = 1 TO NO: PRINT : PRINT "RANGE ";X;: INPUT E(2 * X - 1),E(2 * X): NEXT X: PRINT
     : PR# 1: GOSUB 9999: FOR X = 1 TO NO: PRINT "RANGE #";X;"  ";E(2 * X - 1);"-";E(2 * X): NEXT
     X: PRINT
20760 N = 0: FOR X = 1 TO B%(0): IF I(X) < 0 THEN 20860
20765  FOR Z = 1 TO NO: FOR Y = B%(X) TO B%(X + 1) - 2: IF A(Y) * R4 <  = E(2 * Z) AND A(Y) *
     R4 >  = E(2 * Z - 1) THEN 20840
20820  NEXT Y: GOTO 20860
20840  NEXT Z: PRINT I(X),:N = 1
20860  NEXT X: IF N = 0 THEN  PRINT "NO MOLECULES FOUND"
20870  PRINT : PR# 0: GOTO 20720
20872  REM *
20873  REM *     STATISTICS:  TYPE 3
20874  REM *
20900  PRINT : PRINT "INPUT RANGE";: INPUT R1,R2: IF R1 = 0 AND R2 = 0 THEN 20390
20910  FOR X = 1 TO 15:D(X) = 0: NEXT X: FOR X = 1 TO B%(0): IF I(X) < 0 THEN 20990
20930 N0 = 0: FOR Y = B%(X) TO B%(X + 1) - 2: IF A(Y) * R4 < R1 OR A(Y) * R4 > R2 THEN 20970
20960 N0 = N0 + 1
20970  NEXT Y:D(N0) = D(N0) + 1
20990  NEXT X: PR# 1: GOSUB 9999: PRINT : PRINT "RANGE:  ";R1;" - ";R2: PRINT : PRINT  TAB( 1
     );"BUSHES IN RANGE"; TAB( 20);"# MOLECULES": PRINT :N0 = 0: FOR N = 1 TO 15: IF D(N) = 0
     THEN 21080
21060 N0 = 1: PRINT  TAB( 8);N; TAB( 25);D(N)
21080  NEXT N: IF N0 = 0 THEN  PRINT : PRINT "  ** NO MOLECULES **"
21095  PRINT : PR# 0: GOSUB 9999: GOTO 20900
22000  REM *
22001  REM *     ROUTINE TO PRINT A COPY OF THE CURRENT HISTOGRAM ON THE LINE PRINTER      *
22002  REM *
30000  PR# 1: PRINT "255N";: PRINT  CHR$ (12);: GOSUB 30400
```

```
30030  PRINT  CHR$ (31); CHR$ (3);: FOR X = 1 TO 20: PRINT  CHR$ (3); CHR$ (11);: NEXT X: PRINT
     CHR$ (3); CHR$ (13);:N = 0:I8 =  - 1: GOSUB 30300: PRINT  CHR$ (0); CHR$ (0); CHR$ (0);
     CHR$ (0); CHR$ (15);
30042  FOR X = 1 TO 360 STEP L2: FOR Y = 1 TO L2 - 1: PRINT  CHR$ (1);: NEXT Y: PRINT  CHR$ (
    15);: NEXT X: PRINT  CHR$ (3); CHR$ (13); CHR$ (3); CHR$ (11);
30050  FOR X = 1 TO G9: FOR M = 1 TO G3:N = N + H:I8 = I8 + 1: GOSUB 30300:P = 15: IF I8 < >
    I9 THEN 30130
30100 I8 = 0: PRINT  CHR$ (1); CHR$ (1); CHR$ (1); CHR$ (1); CHR$ (15);: GOTO 30140
30130  PRINT  CHR$ (0); CHR$ (0); CHR$ (0); CHR$ (0); CHR$ (15);
30140  IF H(X) = 0 THEN 30160
30145  FOR Z = 1 TO H(X) * L2: PRINT  CHR$ (P);: NEXT Z
30160  PRINT  CHR$ (3); CHR$ (13); CHR$ (3); CHR$ (11);: NEXT M: NEXT X: GOSUB 30300
30252  PRINT  CHR$ (1); CHR$ (1); CHR$ (1); CHR$ (1); CHR$ (15);
30254  FOR X = 1 TO 360 STEP L2: FOR Y = 1 TO L2 - 1: PRINT  CHR$ (8);: NEXT Y: PRINT  CHR$ (
    15);: NEXT X
30255  PRINT  CHR$ (3); CHR$ (13); CHR$ (3); CHR$ (11);: FOR X = 1 TO 15: PRINT  CHR$ (3); CHR$
    (11);: NEXT X: PRINT  CHR$ (3); CHR$ (2);
30261  PRINT "40N";
30270  PR# 0
30280  RETURN
30290  REM *
30291  REM *     SUBROUTINE TO IDENTIFY HISTOGRAM ON THE PRINTER BEFORE PRINTING IT OUT
30292  REM *
30300  FOR Q = 1 TO 100: PRINT  CHR$ (0);: NEXT Q: RETURN
30400  PRINT  CHR$ (31);: GOSUB 9999: PRINT "FILENAME:  ";C$: PRINT "STEPSIZE= ";G3: PRINT "M
    ULTIPLICATION FACTOR= ";R4: PRINT "BUSHES= ";: IF C1 < > C9 THEN  PRINT C1;"-";
30410  PRINT C9: PRINT "SMOOTHING: ";B$: GOSUB 1000: PRINT "HISTOGRAM COEFFICIENT= ";C: GOSUB
    9999: RETURN
```

A COMPUTER SYSTEM FOR THE ANALYSIS OF DATA GENERATED BY MOLECULAR STUDIES OF DNA

by

Scott McCourt

Submitted to the faculty of the Graduate School in partial
fulfillment of the requirements for the degree
Master of Science in the Department of Computer Science,
Indiana University