

**DESIGN AND IMPLEMENTATION OF A WINDOWS 95/NT**

**GUI FOR GENE MAPPING**

**By**

**LI WANG**

**Bachelor of Science**

**Nanjing Agriculture University**

**Nanjing, China**


**1986**


**Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in Partial fulfillment of  
the requirements for  
the Degree of  
MASTER OF SCIENCE  
July, 1998**

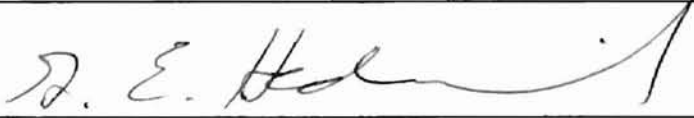
OKLAHOMA STATE UNIVERSITY

DESIGN AND IMPLEMENTATION OF A WINDOWS 95/NT  
GUI FOR GENE MAPPING

Thesis Approved:

  
\_\_\_\_\_  
Thesis Advisor

  
\_\_\_\_\_

  
\_\_\_\_\_

  
\_\_\_\_\_  
Dean of the Graduate College

## ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor Dr. K. M. George for his intelligent, constructive guidance, encouragement, and instruction through my thesis research work. My sincere thanks is also extended to Dr. J. P. Chandler, Dr. G. E. Hedrick, Dr. D. W. Meinke, and Dr. J. E. LaFrance.

I specially thank Dr Meinke and Dr. C. M. Liu for providing me with this research opportunity. Their guidance, support, assistance and friendship were very helpful throughout my study. Without their support and motivation, it would have been difficult to complete this work as it is now.

Finally, I would like to thank my husband and darling son for their love, their understanding, and their encouragement at times of difficulty and sacrifices. Also thanks to all my friends for their support and much needed help.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION .....	1
II. BACKGROUND OVERVIEW .....	6
2.1 Graphical User Interface and Microsoft Visual Basic .....	6
GUI Design Principles .....	7
GUI Principles Applied to GM Project.....	7
2.2 Gene Mapping.....	8
III. GENETIC MAPPING SOFTWARE PACKAGES.....	9
3.1 LINKAGE-1 .....	9
Single-factor Segregation.....	9
Two-factor Segregation.....	10
3.2 JOINMAP .....	10
3.3 MAPMAKER .....	12
3.4 GMENDEL2.0 .....	13
IV. GM PROGRAM OVERVIEW.....	15
4.1 GeneMapping Data File Build Module - BUILD .....	15
4.2 GeneMapping Summary Module - SUMMARY.....	17
4.3 GeneMapping Chi-Square Module - CHI-SQUARE.....	21

4.4 GeneMapping Recombination Estimation Module - RECF <sub>2</sub> .....	24
V. PROGRAM DESIGN AND IMPLEMENTAATION .....	28
5.1 GM Parent Window .....	31
5.2 Startup Window .....	32
5.3 Main Menu Window .....	33
5.4 BUILD Window.....	34
5.5 SUMMARY Window .....	35
5.6 Chi-Square Window.....	37
5.7 RECF <sub>2</sub> Window .....	38
VI. CONCLUSION.....	41
BIBLIOGRAPHY.....	43
APPENDIXES .....	46
APPENDIX A – GENEMAPPING USER’S MANUAL .....	46
Starting and Quitting.....	46
Data Entry and Execution of the Program .....	48
SUMMARY .....	51
CHI-SQUARE .....	54
RECF <sub>2</sub> .....	55
Toolbar and Statusbar .....	57
APPENDIX B – BASIC CONCEPTS OF GENE-MAPPING.....	59

## LIST OF TABLES

Table	page
4.1. F <sub>2</sub> Data File Format.....	17
4.2. SUMMARY Printout.....	20
4.3. AD Test for F <sub>2</sub> Class .....	21
4.4. Four Classes and Their Observed and Expected Frequencies .....	26

## LIST OF FIGURES

Figure	Page
4.1. Flowchart of the SUMMARY Module.....	19
4.2. Chi-Square Flowchart.....	23
4.3. RECF <sub>2</sub> Flowchart.....	27
5.1. MDI Window.....	30
5.2. Relationship Among Windows.....	31
5.3. Parent Window.....	32
5.4. Startup Window.....	32
5.5. Main Menu Window.....	33
5.6. BUILD Window.....	35
5.7. Excel Worksheet.....	35
5.8. SUMMARY Window.....	36
5.9. SUMMARY Print Out.....	36
5.10. Chi-Square Window.....	37
5.11. RECF <sub>2</sub> Window.....	39
5.12. RECF <sub>2</sub> Printout.....	40
A.1. Start Window.....	47

A.2. Exit Dialog Box .....	47
A.3. Main Menu Window .....	48
A.4. Build Window .....	49
A.5. Microsoft Excel Week Sheet.....	49
A.6. Build1 File Window .....	50
A.7. Example for Input File .....	50
A.8. Save as file .....	51
A.9. SUMMARY Window .....	51
A.10. Summary Open File Dialog Box.....	52
A.11. Check the Number of Row and the Number of Column Box .....	52
A.12. Summary Results Screen .....	53
A.13. Summary Printout .....	53
A.14. Check Data File Dialog Box .....	54
A.15. Chi-Square Window.....	54
A.16. Chi-Square Window with Result .....	55
A.17. RECF <sub>2</sub> .....	56
A.18. RECF <sub>2</sub> Window with Result .....	56
A.19. RECF <sub>2</sub> Printout.....	57
A.20. Hide Toolbar .....	58
A.21. Hide Statusbar .....	58



## CHAPTER I

### INTRODUCTION

Genetic maps are a useful tool in various fields of genetic research, both fundamental and applied. The recent developments in molecular genetics, by which large numbers of markers are being generated, have caused a revival of the interest in classic genetic mapping. As a result, linkage analysis and mapping have to a certain extent become computerized. Five computer packages are widely used. The first package is LINKAGE-1 [8]. LINKAGE-1 is a PASCAL computer program designed to aid the geneticist in the detection and estimation of linkage in segregating progenies. Loci segregating for both dominant and co-dominant genes can be simultaneously analyzed. Goodness-of-fit to expected ratios for single-factor segregation's are tested by chi-square analyses. Contingency chi-square analysis is used to test for independent assortment between all pairs of jointly segregating loci. If significant deviation is detected, recombination percentages and their standard errors are calculated.

The second package [22] is Gmendel 2.0. It runs under UNIX operating system and requires a FORTRAN compiler. Gmendel 2.0 generates two-point maximum likelihood estimates for all pairwise matings between all loci by the use of the appropriate segregation class codes and progeny phenotype codes. Linkage phase is correctly assigned based on probability rules, and gene order is estimated using an advanced

multipoint mapping algorithm [22]. Missing progeny data are neither estimated nor substituted and are simply excluded from the two-point estimates.

The third package, called MAPMAKER [10] has been applied to the construction of linkage maps in a number of organisms, including humans and several plants. It uses multipoint linkage analysis for particular types of pedigrees. The program uses an efficient algorithm that allows simultaneous multipoint analysis of any number of loci.

The fourth package is JoinMap [21]. It is a computerized procedure to construct an integrated genetic map. The computer program can handle raw data from  $F_2$  populations, backcrosses, and recombinant inbred lines, as well as listed pairwise recombination frequencies. JoinMap was written in C. Four versions of the executable program are presently available, for PCs (MS-DOS), for VAX systems (VMS), for SUN workstation (UNIX) and for Macintosh computers.

The fifth procedure was written by Koornneef and Stam [9]. It was shown that procedures using  $F_2$  and  $F_3$  populations can be efficiently used for rapid gene localization in a species such as *Arabidopsis* where often double recessive genotypes are not immediately available and where making large numbers of crosses can be difficult. The program runs on IBM PC/AT compatible PCs under MS-DOS.

Several computer packages are presently available for genetic linkage analysis and/or mapping. User friendliness was not a major consideration in the implementation of these packages. A new user may need quite a long learning time to get acquainted with these packages. In this thesis a Gene Mapping (GM) software is developed by the author to assist the geneticist in the analysis of gene localization in a species such as *Arabidopsis* for  $F_2$  data. It is an improved version of a mapping software package

described by Patton [17]. This package is designed for mapping recessive embryo lethal mutations. The previous versions include four computer programs: Build, Summary, Chi-square, and RECF<sub>2</sub>. It is developed as an MS DOS based application.

The BUILD program is used to generate a data file. It prompts users to enter file name, column heading and data. The column heading only allows 3 or 4 columns. If the user makes a mistake in data entry, he or she can correct it only after entering all the data. This is one of the drawbacks of this program.

The SUMMARY program analyzes the data generated by BUILD and totals the plants in each of the four phenotypic classes. This program will output data if the input file is correct.

The CHI -SQUARE program is a computerized version of the Chi-square test. It is used to determine whether the segregation ratios were significantly different from those expected for unlinked genes. The users should enter the data that are output by the SUMMARY program. A menu of choices is provided. One of these choices is to print the results.

The RECF<sub>2</sub> program is a computerized version of the maximum likelihood method. Its main menu includes six items: "choose the phase (coupling/repulsion)," "choose the type of data set," "choose the parameters to estimate," "enter the data (observed frequencies)," "enter the number of iterations," and "enter the trivial value." Each item in this menu must be selected before RECF<sub>2</sub> can make recombination estimates.

The four programs described above form a four step process. The implementation has many drawbacks. First of all, it does not provide a user-friendly approach. Second, the data file can only have 3 or 4 column headings. If the data file is large, then more than one input file should be used, and users may spend a lot of time building the data file. Third, even an experienced user cannot guarantee that he/she will not make any mistakes when creating the data file. If a user enters incorrect data, he/she can correct it only after entering all the data. If the user does not remember where the error lies, the program will provide an incorrect result in the second step. Fourth, the BUILD, SUMMARY, CHI-SQUARE, and RECF<sub>2</sub> are independent programs. The result is that transition is difficult between these programs, e.g., if a user uses the CHI procedure to calculate recombination, he/she must enter data which is a result of SUMMARY, and enter the file name again. The results can not flow automatically from one program to another program. All of these are inconvenient for the user. The user's interest is to obtain satisfactory result for a gene mapping by using a user friendly computer program. However, the above drawbacks pose inconveniences.

The above programs are applications that assist geneticists in gene localization, especially in *Arabidopsis*. As geneticists are users, they would prefer a tool that requires less user interference and has a user-friendly interface.

Microsoft Windows provides a good user interface with its powerful graphical tools. In this thesis, the author developed a user-friendly Graphical User Interface (GUI) for the GeneMapping (GM) program using the Windows environment so that users with even a little knowledge of genetics can obtain results by simply inputting the data

obtained from genetic tests. Therefore the tool can be effectively used not only by experts but also by their staff.

The significant differences between GM and other projects are the following:

- (1) As described above, the GM package provides a user-friendly GUI.
- (2) The project consists of “natural” modules that perform the various tasks in a mapping project.
- (3) The range of the mapping data file is large enough, up to 10 columns, not merely 3 or 4 columns wide.
- (4) The data can be transferred from one module to another module automatically.

The design and implementation will be discussed in detail in chapter V. The language used for the implementation of the project is Visual Basic 4.0. The program runs under Windows 95/NT. To make the description of the project easier to understand, some basic background and concepts of this project are provided in chapter II and in appendix B. Chapter III focuses on past work on gene mapping software. Chapter IV presents the information required by GM (developed in this thesis), including statistical methods, and flow charts of algorithms. Chapter VI summarizes the major improvements of GM software package. Appendix A is a user’s guide for GM software package.

## CHAPTER II

### BACKGROUND OVERVIEW

#### 2.1. Graphical User Interface and Microsoft Visual Basic

Windows software provides an easy-to-use **graphical user interface (GUI)**, with which a user can interact. According to Laudon, the use of a GUI is becoming the trend for today's developers. The command driven and text-based interfaces are no longer the desired form of interface [24]. Command line interfaces are thought to be too complex for users, so a menu-driven interface may be substituted on end user systems [25]. Developers today strive to provide users with an interface that is easier to use and to understand. They want to develop interfaces that are also more efficient and sound, rather than complex, difficult, and unmanageable. According to Leavens [11], understanding 1) what a user interface is and 2) how to build one are the two sides to the user interface problem. Some users may face this problem while trying to create an application that satisfies their needs. All applications written for the Windows environment provide some form of GUI. A graphical user interface that is event-driven would satisfy most users in achieving their tasks. Therefore, a GUI is an attempt to eliminate "typed commands" by allowing users to point at icons on the screen and click with a mouse to direct the computer to perform different tasks. This makes it easy for a beginner to use a computer. In object-oriented/event-driven language, the emphasis of a program is on the objects included in the user interface (such as scroll bars and buttons)

and the events (such as scrolling and clicking) that occur when those objects are used [23].

Visual Basic 4.0 is an object-based/event-driven programming language. It is a powerful Windows application development tool. Almost all of the popular and standard Windows application elements, such as controls and dialog boxes, can be developed in a Visual Basic program.

### **GUI Design Principles**

As Mayhew [13] mentioned, a good GUI application should always keep the user in mind. The user interface is an application's primary mode of communication with the user. Like other forms of communication, whether the developers' ideas will be appreciated depends on how well they are presented. With a good interface design, a program can be efficient and user-friendly. GUI is the key element that distinguishes Windows applications from text-based applications.

### **GUI Principles Applied to the GM Project**

Many experienced GUI developers such as Steve Potts [19] agree that a GUI developer must do a certain level of user analysis and try to balance the needs of different levels of users. This principle is followed in the design of the GM interface. GM has different levels of users. Some users already have a lot of experience with the previous versions. These experienced users may not need many dialog prompts and warning messages, which the new and inexperienced users would need. To help inexperienced users, we must build an application with detailed information presented in the format of both screen dialog boxes and a help system.

## 2.2 Gene Mapping

Evaluation of the genome of human has increased about 45% per year during the past 10 years. If this exponential growth continues, the mapping will include a million STS-based loci (genes) by the year 2005 [26]. The daily updated database of gene markers brings the latest advance of science to the research community. To analyze these databases, more and more computer software packages are being used by geneticists. The application of computer and information technology in biology is expanding to various fields.

The most important application packages in molecular biology have nearly 200 programs, covering all subtopics [27]. Examples are programs for phylogenetic analysis, DNA primer determination, and structural analysis of proteins. GM is based on the project Mapping Genes Essential for Embryo Development in *Arabidopsis thaliana* [17]. This package is designed for mapping recessive embryo lethal mutations.



## CHAPTER III

### GENETIC MAPPING SOFTWARE PACKAGES

The primary genetic linkage maps and procedures for many types of genetic analyses were worked out by geneticists and computer scientists many years ago[9]. These include both plant and natural populations such as humans. About 137 computer software programs on genetic linkage analysis, marker mapping, and pedigree drawing are used by scientists [28]. In this chapter, four gene mapping packages that are related to this thesis are described.

#### LINKAGE-1

LINKAGE-1 [8] is capable of analyzing in a single run an unlimited number of progeny generated from a variety of genetics situations. These families can represent  $F_2$  and backcross types as well as all other combinations of the allowable single -factor segregation ratios. The program accepts both dominant and co-dominant genes as segregating loci. Input for each family consists of single-individual genotype data for each segregating locus.

#### *Single-Factor Segregation*

Goodness-of-fit to expected segregation ratios at each locus is tested by chi-square analyses. The output for each locus consists of the observed segregation and the chi-square and associated  $P$  value for deviations from expected frequencies.

### ***Two-Factor Segregation***

Contingency chi-square is employed to analyze independent assortments between jointly segregating loci. If significance is detected (significance threshold  $P = 0.10$ ), the recombination fraction ( $r$ ) and its standard error (SE) are calculated using maximum likelihood formulae [8].

LINKAGE-1 is written in the PASCAL language and is compatible with the PASCALW compiler. The program is presently dimensioned for a maximum of 15 loci and 300 individuals per progeny and, as such, requires approximately 400K bytes of core memory.

## **JOINMAP**

JoinMap [21] is a computer-implemented procedure to construct integrated genetic maps. Constructing a linkage map is, essentially, the finding of a linear arrangement of markers from recombination values. It runs through the following steps.

1. Read data: The data that can be processed by JoinMap are of several types; examples are raw genotype data from (i)  $F_2$ s, (ii) backcrosses, (iii) recombinant inbred lines (RILs) and (iv) estimates of pairwise recombination percentages together with their standard errors.

2. Calculating pairwise recombination frequencies: First, recombination frequencies are calculated per population. The corresponding LOD values are also calculated. (The LOD score indicates the likelihood of linkage; LOD means the logarithm of odds, the 'odds' being the ratio of the probability that two loci are linked with a given recombination value over the probability that the two are not linked. A LOD value of over 3.0 means that the chances are greater than 1000:1 that the loci are linked for a given recombination estimate. The LOD score decreases with an increasing recombination value; it increases with an increasing sample size, expressing that, e. g. an estimate of 0.3 from a sample of 40 is less informative than an estimate of 0.3 from a sample of 100. LOD values can be seen as a measure of linkage information in the data.). Next, the estimates from distinct populations and the independent estimates (if available) are combined into a single one and the corresponding LOD values are recorded. The estimates of pairwise recombination frequencies from raw data are obtained by maximum likelihood.
3. Establishing linkage groups: The criterion in assigning markers to linkage groups is the LOD value of the pairwise estimates. A threshold LOD value can be set by the user. At any stage in this procedure there is a group of markers which have been assigned to a linkage group and a group of 'free' markers which have not yet been assigned. If none of the 'free' markers is significantly linked to one of the existing groups, a new linkage group is created. Otherwise, the first 'free' marker which does show linkage with an existing group is added to that group.

4. Estimating map distances: The core of the program is the estimation of map distances for a given order of the markers. The algorithm was described by Jensen and Jorgensen [7].

JoinMap was written in C. The CPU time to generate a map with JoinMap depends, apart from hardware configuration, on the internal consistency of the data. Generating a map with 20 markers takes 15-55 CPU-seconds on an 80486 PC, depending on the 'smoothness' of the data.

## MAPMAKER

MAPMAKER [10] was specifically designed for the construction of primary genetic linkage maps from RFLP data either from  $F_2$  intercrosses in experimental populations or from two- and three-generation nuclear families in natural populations. The MAPMAKER program is written in C programming language, with versions for both UNIX and VAX/VMS operating systems.

At the outset of a MAPMAKER session, one loads a file containing either of two types of information, called " $F_2$  data" or "CEPH-type data."  $F_2$  data refers to data from  $F_2$  intercrosses or backcrosses between homozygous inbred lines. The data may contain co-dominant, dominant, or recessive markers, as well as missing data. CEPH-type data refers to data on segregation of co-dominant markers such as RFLPs in two- or three-generation families in a natural population.

The most basic operation in MAPMAKER is to construct the maximum likelihood map for a particular set of loci in a particular order. If one types the command "sequence 9 3 1 7 8," that tells MAPMAKER that the set of five loci numbered 9 3 1 7 8 ,

in this fixed order, should be used in all subsequent analyses. If one next types “map,” MAPMAKER would compute the maximum likelihood map for the five loci numbered 9 3 1 7 8 , in this presumed genetic order. If one types “compare,” MAPMAKER will then compute the maximum likelihood map for each of the orders, will sort the orders by likelihood, and will print out a summary table. One can also type “linked?”. MAPMAKER will ask the user to indicate the interval to test for linkage. The program will then compute (i) the likelihood for the best map if the recombination fraction for the designated interval is held at 50% and (ii) the likelihood for the best map if it is allowed to vary. MAPMAKER provides a number of other commands, including ones that compute the likelihood at any desired point on the likelihood surface.

### **GMENDEL 2.0**

A linkage analysis computer program, Gmendel 2.0 [22] has been developed which allows genome maps to be constructed from any type of diploid cross. GMENDEL 2.0 can perform multipoint linkage analysis on populations with complex genetic structures, such as those arising from an  $F_1$ ,  $F_2$  or backcross between highly heterozygous parents, as well as from more traditional mapping crosses, such as from an  $F_2$  from inbred parents. Any program of general applicability which seeks to create a comprehensive linkage map from heterozygous crosses must be able to generate two-point recombination estimates from all possible matings between multiple segregation types. GMENDEL 2.0 generates two-point maximum likelihood estimates for all pairwise matings between all loci. Linkage phases are correctly assigned based on probability rules and gene order is estimated using an advanced multipoint mapping

algorithm [22]. Missing progeny data are neither estimated nor substituted and are simply excluded from the two-point estimates.

Multipoint gene order is determined by GMENDEL 2.0 using a simulated annealing algorithm (SAA). In brief, it estimates the shortest linear map, the global minimum, by simulating different gene orders for groups of loci in a progressive manner and saving only the shortest orders.

GMENDEL 2.0 runs under a UNIX operating system and requires a FORTRAN compiler.

## CHAPTER IV

### GM PROGRAM OVERVIEW

GeneMapping (GM) is a user-friendly computer software package. This package provides a Windows graphical user interface that can perform gene localization in a species such as *Arabidopsis thaliana*, where one mutant is a recessive lethal. The user interface provides screens for data entry and verification. The user fills in the screen boxes, enters data, and pushes buttons with his/her mouse. GM reads the entered data, collects data, performs genetic and statistical analysis, and displays formatted results.

A user of GM needs a graphics terminal to communicate with GM. Currently, the system also requires a Visual Basic environment. In GM, the various tasks in a mapping project are performed by separate programs (modules).

#### **4.1 *GeneMapping Data File Build Interface – BUILD***

GM uses comma delimited (\*.csv) file to import the data that must be analyzed. A comma delimited file (\*.csv) can be made with the BUILD module. It is the responsibility of the user to prepare a correct data file. All files handled and/or produced by GM are comma delimited files.

An F<sub>2</sub> data file contains information for a single segregating population. It is a sequential file. “Sequential” means that the data are read from left to right, and from top to bottom. The F<sub>2</sub> data file contains column headings at the top of the file. The column

heading names are “the number of pots,” “the number of plants,” the name of visible markers,” and “the embryonic lethal.”

1. *The number of pots.* The F2 data file contains 27 pots. Each pot is marked by number 1-27, respectively.
2. *The number of plants.* Each pot contains 9 plants. Each plant is given a letter (A-I) that corresponds to its position within the pot.
3. *The name of visible markers.* See Table 4.1. in this example, the visible markers' name are “er” “dis.”
4. *The embryonic lethal.* The last column heading should be the embryonic lethal.

The data body contains the actual genotype information for each visible markers and for the embryonic lethal. M indicates that the plant was homozygous for the visible marker or heterozygous for the lethal. T stands for wide type, X stands for a dead plant, B stands for missing data. As described in Chapter I, these data files allow any number of rows and any number of columns up to 10. There is no space between the rows. A file name can be up to 32 characters with no spaces.



Table 4.1 F<sub>2</sub> Data File Format

Pot #	Plant #	er	dis1	emb30
1	A	X	X	X
1	B	T	T	M
1	C	T	M	T
1	D	M	T	M
1	E	T	T	M
1	F	X	X	X
1	G	M	M	M
1	H	T	T	T
1	I	M	T	M
2	A	T	M	T
2	B	M	T	M
2	C	M	T	M
2	D	M	M	M
2	E	X	X	M
2	F	T	M	X
2	G	T	T	T
2	H	M	B	M
2	I	T	T	T

#### 4.2. GeneMapping Summary Module – SUMMARY

The first step in the data analysis of the mapping project is to analyze the BUILD data and the total plants that fall into each of the four phenotypic classes (A, B, C, D) as described elsewhere in this thesis. This task is performed by the SUMMARY module.

The SUMMARY module compares the marker's phenotype with the embryo mutant ( the last column in Table 4.1). If both phenotypes are "+", then the plant is grouped into the "A" class. If the marker's phenotype is "+" and the embryo mutant phenotype is "M", the plant is grouped into the "B" class. If the marker's phenotype is "M" and the embryo mutant phenotype is "+", then the plant is grouped into the "C" class. Finally, if both the

marker phenotype and the embryo mutant are “M”, the plant is grouped into the “D” class. The total number of dead plants can also be summarized. The flow chart shown in Figure 4.1 provides details of the SUMMARY program. An example of the SUMMARY printout is shown in Table 4.2.

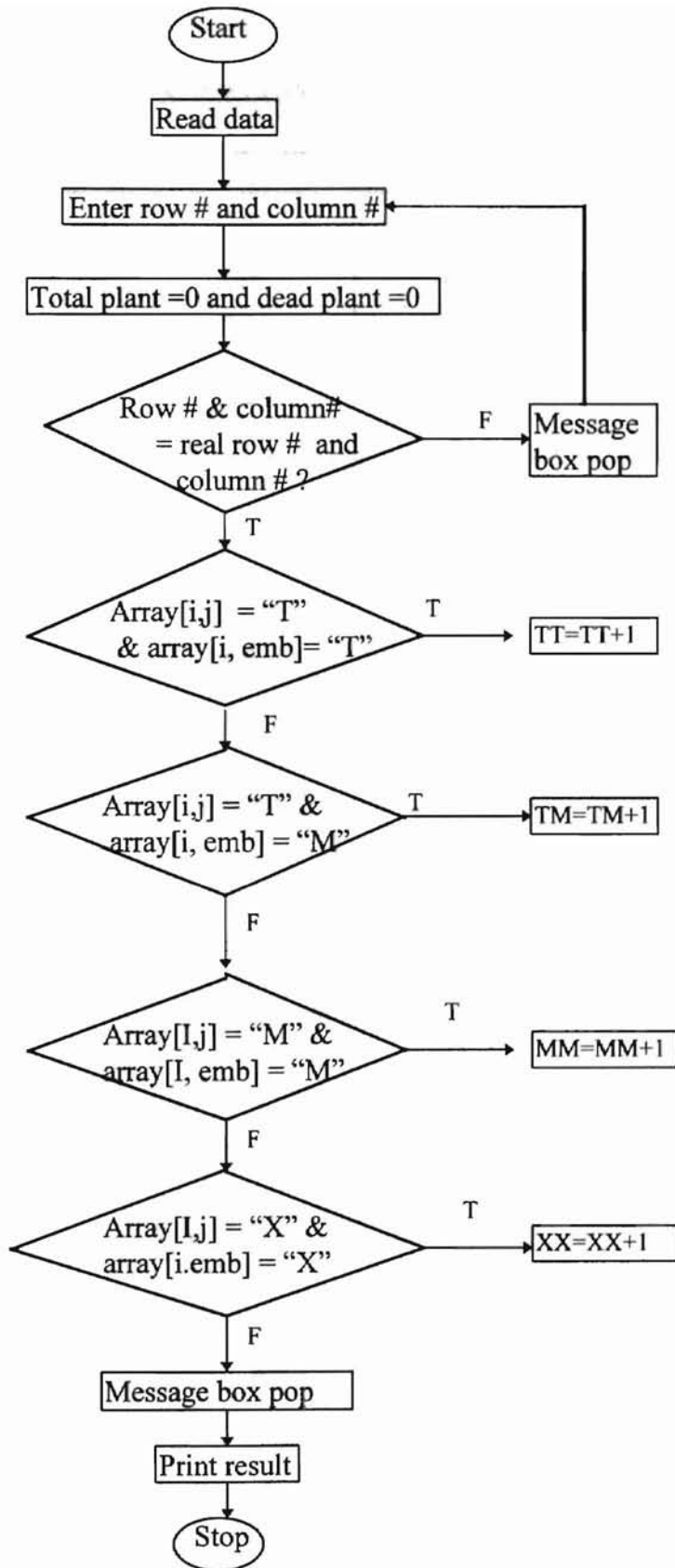


Figure 4 .1 Flowchart of the SUMMARY Module

Table 4.2 SUMMARY Printout

Pot #	Plant #	Dis1	CIV2	EMB127
SUMS OF:				
A	++	12	62	
B	+M	127	116	
C	M+	62	12	
D	MM	16	27	
TOTAL NUMBER OF PLANTS:		243		
TOTAL NUMBER OF DEAD PLANTS:		26		

Another test of the SUMMARY module is to check if the input data is correct. If the user enters incorrect data, SUMMARY will show a message box containing a warning to check the data file. For instance, valid input consists of the characters M, X, T, and B. If a user inputs "w" from the keyboard, the message box will show "incorrect character, please check input data file!"

The SUMMARY program depends on two important parameters, the number of rows and the number of columns. The number of rows includes table heading or the total number of plants plus one. For example, if the number of pots is 27, and there are 9 plants per pot, then  $row\# = 27 * 9 + 1 = 244$ . The number of columns is the sum of all visible markers count plus 3. For example, if there are 5 visible markers, the  $column\# = 5 + 3 = 8$ . Let us also use the previous example, Table 4.1. In the previous example, the total number of plants is 14, so the number of rows is  $14 + 1 = 15$ . The visible markers are "er" and "dis1," so the number of columns is  $2 + 3 = 5$ . Both the number of rows and the number of columns are required input for SUMMARY interface row and column boxes. If the row count's and column count's input are greater than the actual row and column

counts of the file, GM will try to read beyond the end of the file, which will also lead to an error message. When row count and column count are smaller than the actual row and column count, it will issue a warning indicating there are more data in the file.

#### 4.3 GeneMapping Chi-square Module – CHI-SQUARE

Chi-square method ( $\chi^2$ ) is a statistical procedure that enables the investigator to determine how closely an experimentally obtained set of values fits a given theoretical expectation [4] [6]. The degrees of freedom equals the total number of pairwise estimates available for the current map minus the number of adjacent intervals. In the AD test (A:B:C:D= 3:6:1:2 ratios), there are 3 degrees of freedom [18], see Table 4.3. A formula for  $\chi^2$ , designed for a sample consisting of four classes (A:B:C:D) is symbolized as follows

$$\chi^2 = \sum \frac{d^2}{e}$$

where  $d$  is the deviation between each observed and expected class value,  $e$  is the expected value in the respective class, and the  $\Sigma$  is the summation sign. As an example,  $\chi^2$  is calculated for the AD method (Table 4.3).

Table 4.3 AD Test for F<sub>2</sub> Class

	F <sub>2</sub> class			
	A	B	C	D
Observed	108	252	54	53
Expected	116.75	233.50	38.76	77.99

$$\chi^2 = \frac{(108 - 116.75)^2}{116.75} + \frac{(252 - 233.5)^2}{233.5} + \frac{(54 - 38.76)^2}{38.76} + \frac{(53 - 77.99)^2}{77.99} = 16.123$$

The next step is to interpret the  $\chi^2$  value in terms of probability. The percent point ( $P=5\%$ ) is usually chosen as an arbitrary standard for determining the significance or goodness of fit. In general, let  $P$  represent the probability of obtaining a deviation as great as or greater than that obtained from the experiment by chance alone. If  $P$  is small, it is concluded that the deviations are not due entirely to chance, and the hypothesis is rejected. If  $P$  is greater than the predetermined level ( $P=0.05$ ), the data conform well enough and the hypothesis is accepted [17]. In this thesis, the level of statistical significance is defined as follows: \* indicates there is a significant difference between expected value and observed value at  $P = 0.05$ , \*\* at  $P=0.01$ , and \*\*\* at  $P \leq 0.005$ , respectively. The  $\chi^2$  algorithm is shown as Figure 4.2.

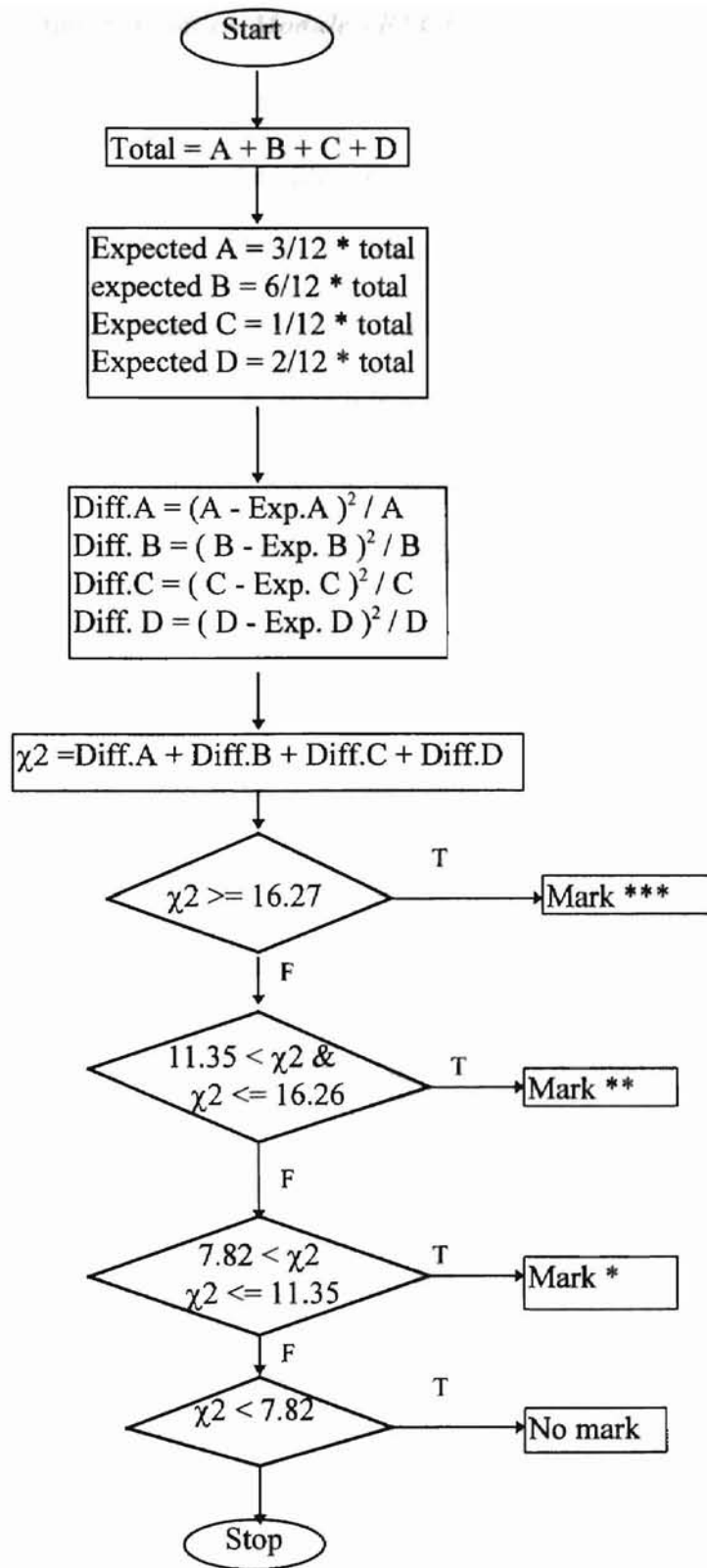


Figure 4.2  $\chi^2$  Flow Chart

#### 4.4 Genemapping Recombination Estimation Module - RECF<sub>2</sub>

The RECF<sub>2</sub> module performs one of the core tasks in a mapping project: it estimates recombination frequencies from F<sub>2</sub> data file, as described in Chapter I.

Recombination estimates are obtained by maximum likelihood.

Fisher's method of maximum likelihood [14] [1] is a well-known technique for deriving estimates of unknown parameters. It is often impossible to obtain explicit solutions of the maximum-likelihood equations, especially if several parameters are involved, and iterative methods of numerical approximation are required. A widely used technique is the one normally referred to as maximum-likelihood scoring. Let us suppose that for a sample of  $n$  observations all independently distributed with the same frequency function, the logarithm of the likelihood is  $L$ . We also suppose that  $L$  is a function of just one unknown parameter  $\theta$ . Then the usual maximum-likelihood equation for  $\theta$  is

$$S(\theta) \equiv \frac{dL}{d\theta} = 0 \quad (4.1)$$

with solution  $\hat{\theta}$ . The expected amount of information in the sample is defined as

$$I(\theta) = -E \frac{d^2 L}{d\theta^2} \quad (4.2)$$

where the expectation  $E$  is taken over all samples of size  $n$ , and in sufficiently large samples the variance of  $\hat{\theta}$  is  $\{I(\theta)\}^{-1}$ .

The score for  $\theta$  is simply the function  $S(\theta)$  defined in (4.1). One way of solving this equation is to choose a trial value  $\theta_I$  which makes  $S(\theta_I)$  fairly small. Thus  $\theta_I$  is a first approximation to  $\hat{\theta}$ . Then calculate  $I(\theta)$  and obtain the improved estimate



$$\theta_2 = \theta_1 + S(\theta_1)/I(\theta_1) \quad (4.3)$$

In some case, it is more convenient to calculate the score at a few suitable trial values and so to avoid the direct use of (4.2).

Assume that

$$S(\theta_2) < 0 < S(\theta_1), \quad (4.4)$$

where  $S(\theta_2)$  and  $S(\theta_1)$  are both small. then  $\theta_1 < \theta_2$  since the  $S$  must be a decreasing function of  $\theta$ . Then we can interpolate linearly between  $\theta_1$  and  $\theta_2$  by the 'rule of false position' to give the maximum-likelihood estimate

$$\hat{\theta} \approx \frac{-\theta_2 S(\theta_2) + \theta_1 S(\theta_1)}{S(\theta_1) - S(\theta_2)} \quad (4.5)$$

while the observed amount of information is estimated by

$$I \approx \frac{S(\theta_1) - S(\theta_2)}{\theta_2 - \theta_1} \quad (4.6)$$

suppose that the  $n$  observations fall into  $k$  classes, of which the  $i$ th contains  $a_i$  individuals with expectation  $m_i(\theta)$ . We take logarithms and differentiate with respect to  $\theta$ , which gives the score

$$S(\theta) = \sum_{i=1}^k \frac{a_i}{m_i} \frac{dm_i}{d\theta} \quad (4.7)$$

while the amount of information is

$$I(\theta) = \sum_{i=1}^k \frac{1}{m_i} \left( \frac{dm_i}{d\theta} \right)^2 \quad (4.8)$$

the standard error of expectation is

$$\text{stdp} = 1/\sqrt{I(\theta)} \quad (4.9)$$

The RECF<sub>2</sub> flowchart is shown in Figure 4.3. The following simple example illustrates the idea. Let four classes and their observed and expected frequencies be as given in Table 4.4 [5].

Table 4.4 Four Classes and Their Observed and Expected Frequencies

Classes	A	B	C	D
Observed	51	108	11	49
expected frequencies	$n(2p-p^2)/3$	$2n(1-p+p^2)/3$	$n(1-p)^2/3$	$2np(1-p)/3$
Expected ratio	3	6	1	2

By means of the method of maximum-likelihood estimates, it is possible to choose as the estimate of the unknown parameter  $p$ , the value that maximizes the probability of the observed numbers. If we choose  $p=0.25$ , the maximum-likelihood equation for  $p$  is

$$sp=(a*(2-2*p))/(2*p-p^2) + (-b+2*b*p)/(1-p+p^2) + (-2*c+2*c*p)/(1-2*p+p^2)+(d-2*d*p)/(p-p^2),$$

the expected amount of information is

$$Ip=2/3*n*(2*(1-p)^2/(p*(2-p) + (2*p-1)^2/(1-p+p^2) + 2*(p-1)^2/(1-p)^2+ (1-2*p)^2/(p-p*p)$$

with the standard error

$$\text{stdp}=1/\text{sqr}(Ip).$$

GM also calculates the iteration number.

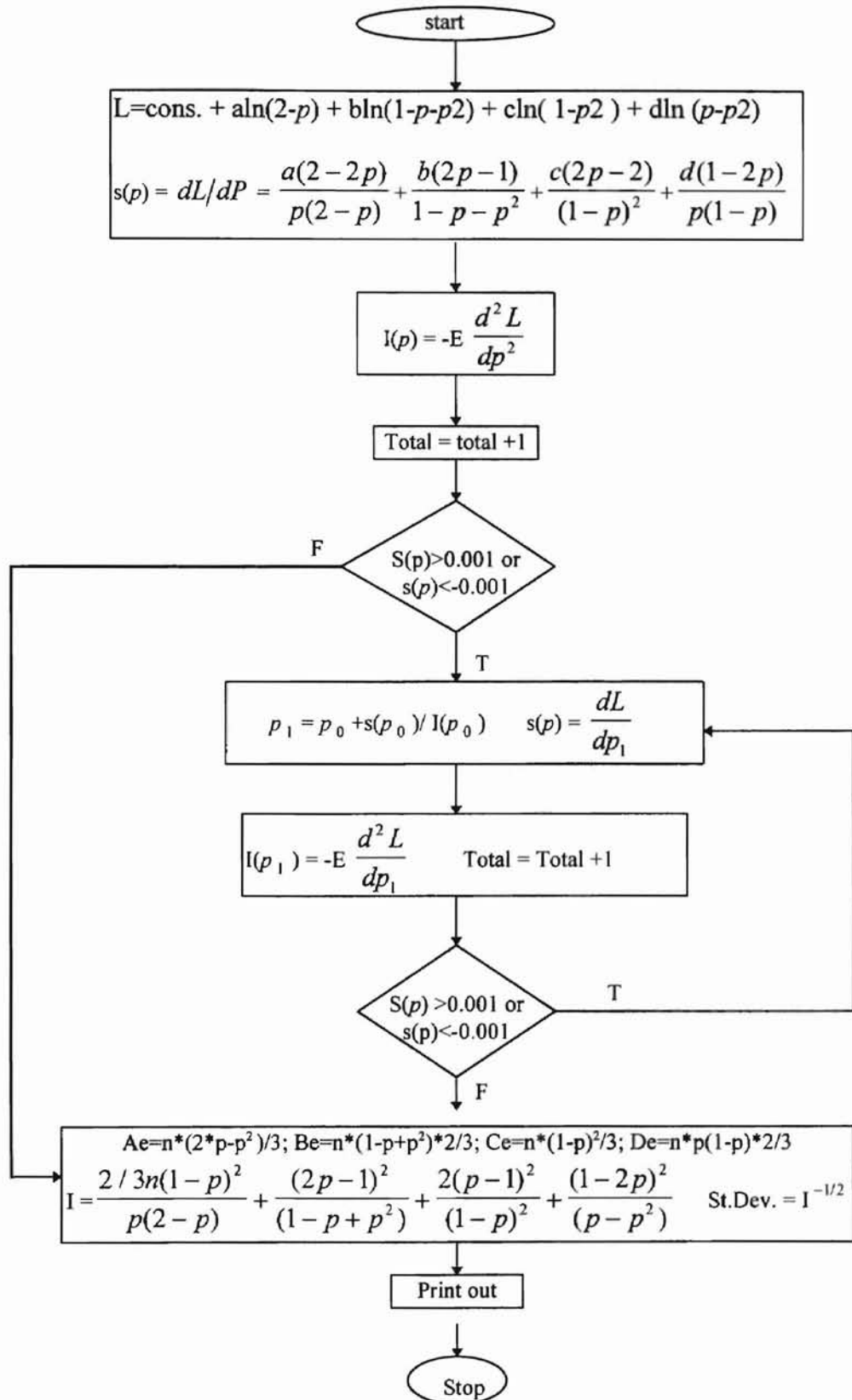


Figure 4.3 RECF<sub>2</sub> Flowchart

o

**CHAPTER V**  
**PROGRAM DESIGN AND IMPLEMENTATION**

Windows GUI design was one of the most important parts of GM implementation. The Graphical User Interface (GUI) is implemented as a Multiple Document Interface (MDI) window which includes a tool bar, a scroll bar which separates the data entry and graphical plot of data, dialog boxes, control button, etc. It provides a convenient way for users to select a desired dialog box for data entry or to execute a desired command.

Microsoft Visual Basic 4.0 provides many controls for building Windows applications. The author used seven different types of controls in this project. Some of the major controls are listed below:

- Text box control: provides an area in the form where the user can enter data.
- Label control: displays text not to be modify by the user, i.e. all the data field names are labeled by using the label control.
- Frame control: groups data in the same screen into different frames according to control type.
- Combobox control: this control lets the user select data instead of directly inputting data.

- Option button control : limits the user to only one of two or more related and mutually exclusive choices. In the main menu, the option button control is used to allow the user to select an option.
- Timer control : is used to process code at regular time intervals.
- Command button: performs an immediate action when clicked. Most actions in GM are performed by using the command button.

As a Multiple Document Interface (MDI), GM contains seven documents (Figure 5.1). They are the parent window, startup window, main menu window, build window, summary window, chi-square window, and RECF<sub>2</sub> window. Each document has its own window, and all the documents are contained within a single parent form. Every window contains a set of buttons. The different buttons lead programs to different windows. For instance, main menu window contains BUILD, SUMMARY, CHI\_SQUARE, RECF<sub>2</sub>, and OPEN buttons. When users click BUILD button, the program goes to BUILD window. The OPEN button lets program go to SUMMARY window. Figure 5.2 shows the relationship between these windows. The windows are described in the following sections.

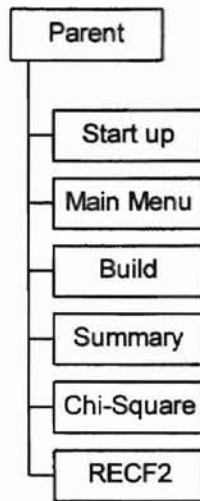


Figure 5.1 MDI Windows

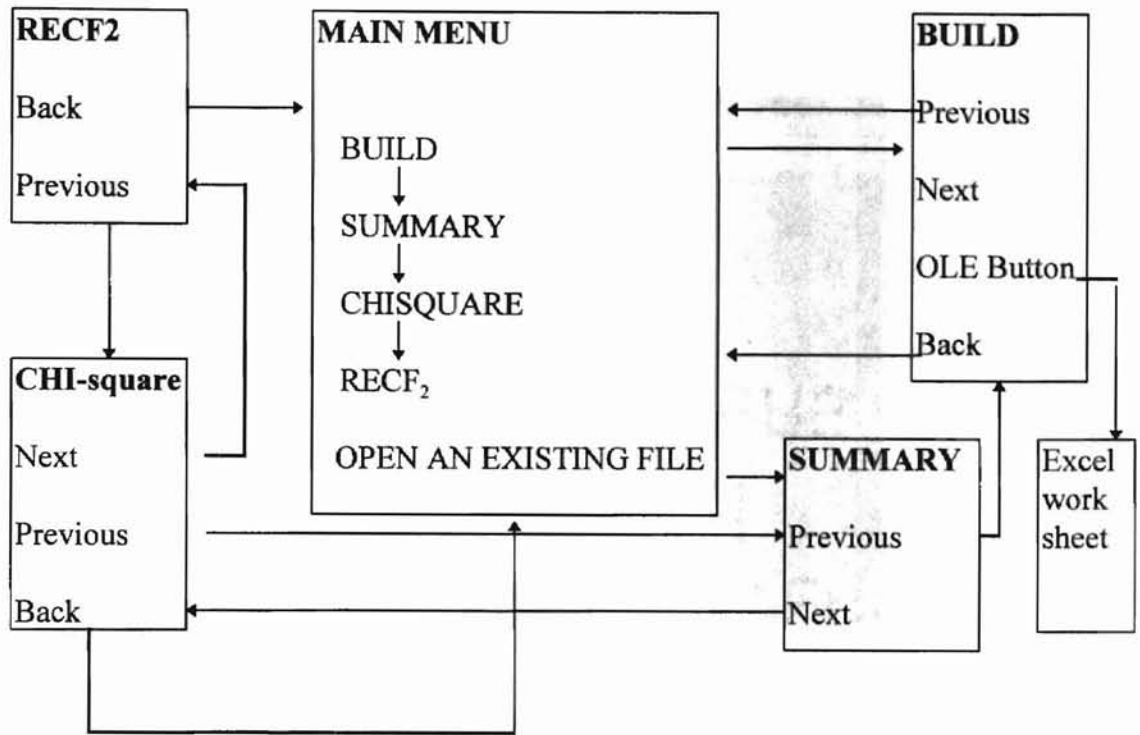


Figure 5.2 Relationship among Windows

**5.1 GM Parent Window:** As Figure 5.3 shows, the GM parent window contains five separate parts: title bar, menu bar, toolbar, status bar and form window. The title bar shows the project name -- Gene Mapping. It also contains the minimize button, maximize/resize button, and close button. Below the title bar is the menu bar. The menu bar displays the commands that will be used to build GM applications, i.e. File, Windows, View, and About. The buttons on the toolbar located below the menu bar, provide quick access to the most commonly used menu commands. The form window is a window in which the children such as build, summary, chi-square, and RECF<sub>2</sub> windows should be located.

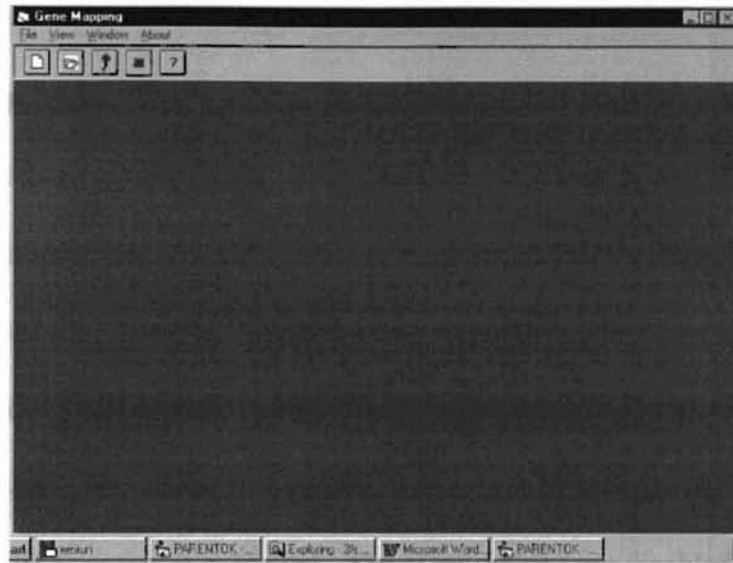


Figure 5.3 Parent Window

**5.2 Start up Window:** The start up window (Figure 5.4) is designed as a cover for the GM project. It contains a picture of the DNA double helix. This window also includes the title of this project, that is "Gene Mapping for Windows." When GM begins execution, the startup window stays in the parent form only 3 seconds. This is controlled by using the timer control.

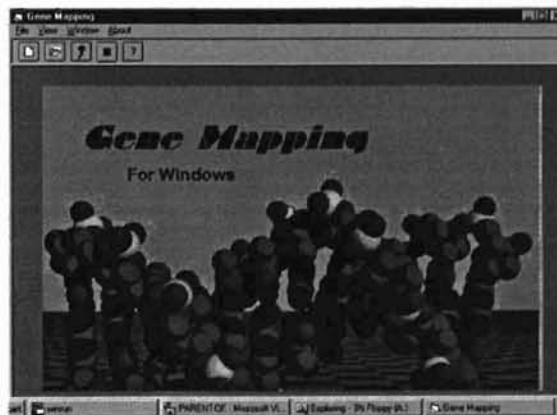


Figure 5.4. Startup Window



5.3 **Main Menu Window:** The main menu window contains a set of options that are grouped into three separate parts. Each part is in a different frame. The main menu can be used to either create new F<sub>2</sub> data file or to open an existing data file. The first part is used to create new F<sub>2</sub> data file. When the user wants to create a new F<sub>2</sub> data file, he/she must first use the BUILD module to build the data file, then follow the instructions to perform a set of analyses. In order to prevent a new user from choosing the wrong selection, the BUILD option is activated. The other options are disabled. When the Build option and the OK command are clicked, the corresponding build window opens and displays a set of options that allow the user to go to the next operation. The second part is “Open an Existing File.” If the data file has been prepared, this option can be selected. When the user clicks this option, the SUMMARY window pops up. The third part is a set of command buttons. They are OK, Help, and Exit buttons as shown in Figure 5.5.

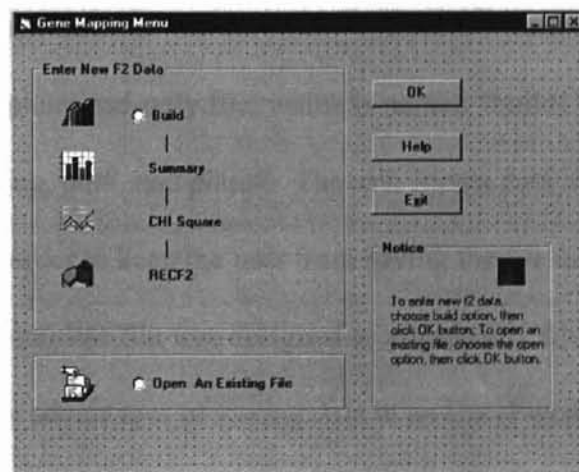


Figure 5.5 Main Menu Window

**5.4 Build Window:** The Build window is designed for the user to build data files. To analyze the big database for markers identified in Meinke's laboratory, it is necessary for GM to exchange information. Microsoft released a new standard known as object linking and embedding (OLE), which lets users share information between almost any pair of applications running on the windows desktop. OLE lets users embed information within other programs, like an Excel spreadsheet. The data that is copied from one application to another is referred to as an object. The application that created the data (object) is called the server application, and the application to which the object is copied is called the client application. OLE allows users to start a Windows application from inside another Windows application. For example, OLE enables users to start Excel from a Visual Basic application that contains an Excel work sheet [23]. In the BUILD windows, the author, using OLE, linked the object to the client application, which is the visual Basic application—GM. When the user double clicks on the spreadsheet icon, (Figure 5.6), the Microsoft Excel worksheet pops up (Figure 5.7). In order to standardize the format of the input file and make it convenient to user, the author designed a template read-only file, which is named "build1". This template file contains column heading, pot#, and plant#. The user inputs data in the corresponding rows and columns. In order to keep the user from saving the file under the same name as the default name, the template file was designed as a read-only file.

There are many advantages of storing data in an Excel worksheet. Data stored in that format can be retrieved both quickly and easily by the computer. In addition, using a template is very convenient for the users.

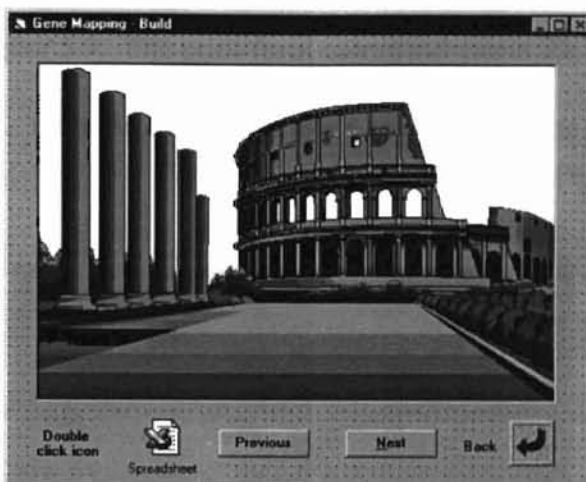


Figure 5.6 Build Window

Pot #	Plant #	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9	ch10	ch11
1	1A	l	m	l	l	l	l	l	l	l	l	l
2	1B	x	x	x	x	x	x	x	x	x	x	x
3	1C	x	x	x	x	x	x	x	x	x	x	x
4	1D	x	x	x	x	x	x	x	x	x	x	x
5	1E	l	l	l	l	l	l	l	l	l	l	l
6	1F	l	l	l	l	l	l	l	l	l	l	l
7	1G	x	x	x	x	x	x	x	x	x	x	x
8	1H	x	x	x	x	x	x	x	x	x	x	x
9	1I	l	l	l	l	l	l	l	l	l	l	l
10	2A	l	l	l	l	l	l	l	l	l	l	l
11	2B	l	l	l	l	l	l	l	l	l	l	l
12	2C	m	l	l	l	l	l	l	l	l	l	l
13	2D	l	l	l	l	l	l	l	l	l	l	l
14	2E	x	x	x	x	x	x	x	x	x	x	x
15	2F	x	x	x	x	x	x	x	x	x	x	x
16	2G	x	x	x	x	x	x	x	x	x	x	x
17	2H	l	l	l	l	l	l	l	l	l	l	l
18	2I	l	l	l	l	l	l	l	l	l	l	l
19	3A	l	l	l	l	l	l	l	l	l	l	l
20	3B	x	x	x	x	x	x	x	x	x	x	x
21	3C	x	x	x	x	x	x	x	x	x	x	x
22												

Figure 5.7 Excel Worksheet

**5.5 Summary Window:** The SUMMARY window (Figure 5.8) is designed to summarize the data from a file which is built in the Build module. After the user clicks the “open” icon, the dialog box allows the user to select the data file that will be analyzed. The column and row text boxes allow the users input the number of columns and the number of rows from the keyboard. If the user inputs the wrong column number and row number corresponding to the file, the dialog box will pop up to warn the user to reenter the data and terminate the summary processor. The print command prints the

results as shown in Figure 5.9. The print-form command prints the SUMMARY window as shown in Figure 5.8. Because of Dr. Meinke's requirement, the result form has 10 columns

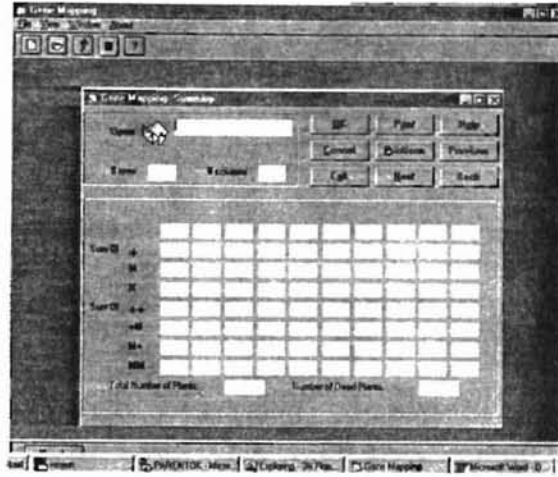


Figure 5.8 SUMMARY Window

		Title A:\test.csv					
		ch1	er	g11	cer2	tt3	emb
sum of	+	139	113	119	119	133	78
	M	19	45	39	39	25	80
	other	85	85	85	85	85	85
Sum of:							
	++	64	52	58	58	66	78
	+M	75	61	61	61	67	0
	M+	14	26	20	20	12	0
	MM	5	19	19	19	13	80
Total Number of Plants:		243					
Number of Dead Plants:		85					

Figure 5.9 SUMMARY Print Out

**5.6 Chi-Square Window:** The chi-square window (Figure 5.10) is designed to perform statistical analysis using the algorithm as shown in Figure 4.2. The title box shows multiple marker lines (DP23, DP24, DP28) crossed with the visible markers (ch1, tt3, etc.). For example, Combobox dropdown list displays a list of multiple marker lines (DP23, DP24, DP28 ) and a list of visible markers (ch1, er, etc.). Users can choose visible markers instead of entering marker names into the combo box. If the user chooses DP23 and the chi visible marker, the title should be “DP23\*ch1.” When the user clicks the display command, the display command displays the SUMMARY results corresponding to the visible marker into the chi-square window. The A, B, C, D (A representing (++) , B representing (+M), C representing (M+), and D representing (MM)) corresponding to the observed data appear. If the user clicks the OK button, the result shows on the screen. The “Print-Form” button prints form as shown in Figure 5.10. The back main menu icon lets users return to the main menu.



Figure 5.10 Chi Square Window

**5.7 RECF<sub>2</sub> Window :** RECF<sub>2</sub> window (Figure 5.11) is designed to perform the core process in gene mapping. The algorithm is shown in Figure 4.3. It contains the following controls.

1. *Title box:* Shows the same visible markers as the Chi-Square window.
2. *Trial value recombination box:* The trial value is used for estimating the level of the recombination. The program uses this value as a starting point as it tries to fit data with what would be expected at different levels of recombination. Usually, the trial value default is 0.25 ( 25% recombination). The user can also enter a higher or lower value instead of 0.25, depending upon the data.
3. *Iteration box:* Shows the number of iterations which seem to work well as long as segregation data are fairly close to what is expected for some percent recombination.
4. *Observed box:* Shows observed A, B, C, D values corresponding to the title (visible marker).
5. *Estimate box:* Shows the recombination estimate value by using the maximum-likelihood algorithm.
6. *St.Dev. box:* Shows the standard deviation value.
7. *Display control:* Displays A, B, C, D observed values and title when the user enters this button.
8. *OK visual basic control:* When the user clicks this button, the expected values and chi-square contribution will display on the screen. The total chi-square

value is used to compare the observed data to what is expected for that level of recombination.

9. *Print control*: Prints the results as shown in Figure 5.12.
10. *Previous control*: Allows the user to return to the chi-square window to perform another visible marker analysis.
11. *Printform control*: Prints the form as shown in Figure 5.11.
12. *Back icon*: Allows the user to go back to the main menu.

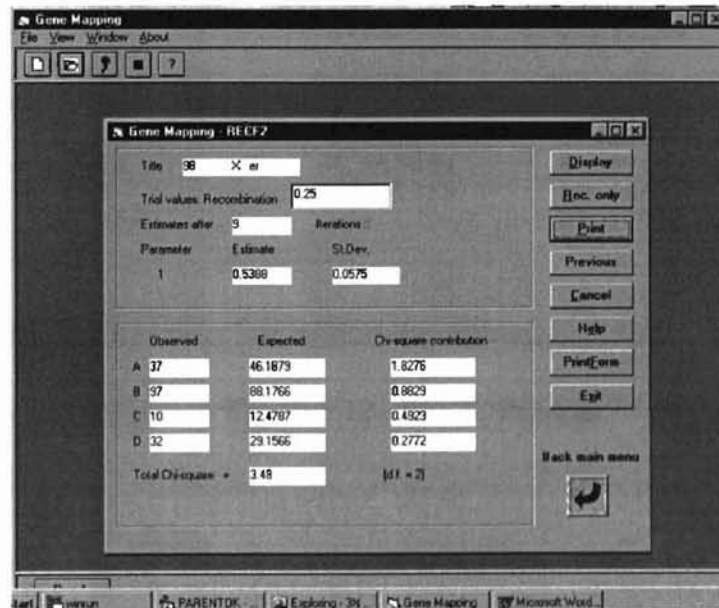


Figure 5.11. RECF<sub>2</sub> Window

	DP23	X	chl	
	Observed	Expected		Difference
	64	39.5		24.5
	75	79		-4
	14	13.16667		0.8333
	5	26.33333		-21.3334
Total	158	158		

CHI\_square = 32.73418 \*\*\*

No. of calculation= 7

Parameter	Estimate	St.Dev.
1	0.62633	0.05951045

Observed	Expected	Chi-square Contribution
64	45.313	7.706482
75	80.68114	0.4000355
14	7.353664	6.007043
5	24.65219	15.6663
Total	158	29.77986 (d.f.=2)

Figure 5.12. RECF<sub>2</sub> Printout



## CHAPTER VI

### CONCLUSION

An object-oriented / event-driven application of GeneMapping (GM ) has been presented, which is based on the project “Mapping Genes Essential for Embryo Development in *Arabidopsis thaliana*” that has been supported by National Science Foundation & the Samuel Roberts Noble Foundation since 1988. GM is a new version of an existing gene mapping software. The major contribution of GM is a Graphical User Interface (GUI) with features of a pull-down menu, toolbar, droplist, control button, message box windows and object linking and embedding, running on Windows 95/NT operating system. The object -oriented/event-driven language Visual Basic has been employed to implement genetic and statistical analysis and the graphical user interface of the project. Various tests have been performed by different modules.

There are several advantages of using the GM project.

First, users don't have to be genetic experts in order to analyze collected data from genetic tests when following the user's guide. Thus, more time and attention can be paid to genetic tests rather than human errors while calculating the statistical rests.

Second, GM is user-friendly with interactive communication with users. Users are able to select options from a main menu, by continuing to click the command button to get exact statistical results.

Third, the OLE technique has been employed to design the Build module, which helped to develop a flexible data file module. That enables user to have any size for the input file. In addition, the template read-only file model is used to construct data files. The users can easily and conveniently build data files. These are the main differences from the original package that are worth highlighting.

Finally, the GM package has been split into modules. Data translation between different modules is performed by global variables. Users can simply click the display button instead of entering the previous module's results manually, which is more convenient for the user.

## BIBLIOGRAPHY

1. Allard, R. W. "Formulas and Tables to Facilitate the Calculation of Recombination Values in Heredity." Hilgardia, 24, 235-278,1956.
2. Brisco, R.M. A Multiple Windows Interface for Internet Tools. Langston University,1996
3. Eckel, D. G., Montante, J. C., & Garcia, K. PrintAPlot Emeryville: Insight Development Corporation, 1988
4. Gardner, J.E., Principles of Genetics. New York: Wiley,1968.
5. Franzmann, H. L., Yoon, E.S., & Meinke, D.W., Saturating the Genetic map of *Arabidopsis Thaliana* with Embryonic Mutations. The Plant Journal Vol. 7(2), 341-350.1995.
6. Fisher, R. A., The Genetical Theory of Natural Selection. New York: Dover Press, 1958.
7. Jensen, J., Jonrensen, J.H., "The barley chromosome 5 linkage map." Hereditas, Vol. 80, 5-16, 1975.
8. Suiter,K.A., Wendel, J.F., & Case, J.S., "LINKAGE - 1: a PASCAL computer program for the Detection and Analysis of Genetic Linkage." The Journal of Heredity, Vol. 74, 203 - 204, 1983.
9. Koornneef, M., Stam, P., "Procedures for Mapping by Using F2 and F3 Populations." Arabidopsis Infor. Serv., 25, 35-40, 1984.

10. Lander, S. E., Green, J.P., Barlow, A. A., Daly, J. M., Lincoln E.S., & Newburg, L.,  
“MAPMARK: An Interactive Computer Package for Constructing Primary Genetic  
linkage Maps of Experimental and Natural Populations.” Genomics, Vol. 1, 174-181,  
1987.
11. Leavens, A., Designing GUI Applications for Windows, M&T Book, A Division of  
MIS: Press, Inc., A Subsidiary of Henry Holt and Company, Inc.1994.
12. Mann, T., Real-world Programming with Visual Basic. Indianapolis, Ind. Sams pub.  
1995.
13. Mayhew, D. J., Principles and Guidelines in Software User Interface  
Design.Englewood Cliffs:P T R Prentice Hall,1992.
14. Bailey, Bailey, N.T., Introduction to the Mathematical Theory of Genetic Linkage.  
Oxford: Clarendon Press, 1961.
15. Nathan & Ori Gurewich, O., (1993). Teach Yourself Visual Basic in 21 Days. Sams  
Publishing
16. King, C.R., A Dictionary of Genetics. New York: Oxford University Press, 1974
17. Patton, D.A., Mapping Gene Essential for Embryo Development in *Arabidopsis  
thaliana*. Stillwater: Oklahoma State University, 1991.
18. Patton, D.A., Franzman, L.H., & Meinke, W.D., “Mapping Gene Essential for  
Embryo Development in *Arabidopsis thaliana*. “Mol Gen Genet , Vol. 227, 337-347,  
1991.
19. Potts,S., Multiple Document Interface. Visual Basic 4 Expert Solutions. Que  
Corporation, Indianapolis, In. 1996.

20. Servitova, J., Cetl, I., "The use of Recessive Lethal Chlorophyll mutants for linkage mapping of *Arabidopsis thaliana* (L.) Heynh." Arabidopsis Inform. Serv., Vol. 21, 59-64, 1984.
21. Stam, P., "Construction of Integrated Genetic Linkage Maps by means of a New Computer Package: JoinMap." The Plant Journal, Vol. 3(5), 739-744, 1993.
22. Knapp, S., Liu, B.H., "Genome Mapping with Non-inbred Crosses Using Gmendel 2.0." Maize Genetics Cooperation News Letter, 27 - 29. 1992.
23. Zak, D., (1997). Programming with Microsoft Visual Basic 4.0 for Windows. Cambridge: ITP Inc. 1997.
24. Laudon, Kenneth, C., & Jane, P., Essentials of Management Information Systems. Prentice Hall, Inc., 1995.
25. Tenopir, Carol, "The User-system Interface," Library Journal, Vol. 114(13), PP.88-81, 1989.
26. Collins, A., Frezal, J., Teague, J., & Morton, N.E., "A metric Map of Humans: 23,500 Loci in 850 Bands." Medical Sciences. Vol. 93, pp. 14771-14775, Dec. 1996.
27. Web site: <http://www.seitti.funet.fi/english/biosciences.html>.
28. Web site: <http://www.linkage.rockefeller.edu/soft/list.html>.

## APPENDIX A

### GENEMAPPING USER'S MANUAL

Gene Mapping is an object-oriented /event-driven approach to estimating gene location in a species such as *Arabidopsis thaliana*, where one mutant is a recessive lethal.

GM is an MDI windows application, providing a Graphical User Interface ( GUI ) environment in terms of pull-down menus, dialog boxes, and drop lists, etc. Users with limited knowledge of genetics are able to perform gene location analysis by simply inputting the data resulting from genetic tests. The requirements to run GM project is Windows 95 or Windows NT.

#### Starting and Quitting

There are four ways to run GM in Windows 95/NT.

(1) In the File Manager, select A drive, find GM.EXE and double click on it.

GM starts and displays the DNA double helix start window, then displays the main menu window.

(2) In the status bar, from the start window, choose RUN, type A:\GM.EXE on the command line and click the OK button.

(3) In the program windows explorer, choose A drive, find GM.EXE and double click on it. GM starts and displays the DNA double helix window.

(4) Double click on Microsoft Visual Basic. From the File menu, choose

open and select A drive, GM file; click OK. Then click the RUN button in the toolbar.

The first three ways require no source files, no Microsoft Visual Basic tool, only the executable file GM.EXE. The fourth way of starting GM needs all the source files of the GM and Microsoft Visual Basic environment. No matter which way you select to use, the GM windows application displays on the screen as shown in Figure A-1.

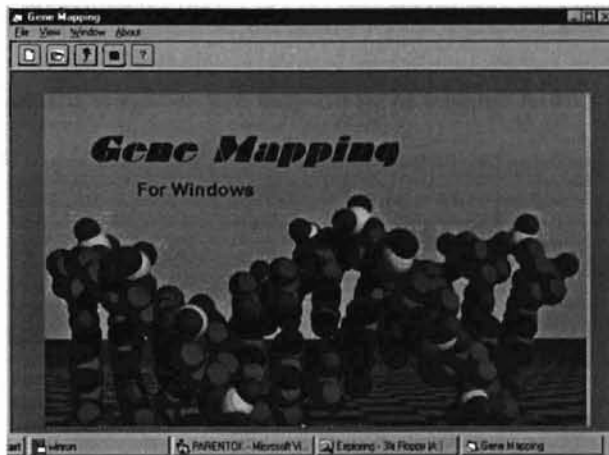


Figure A-1 Start Window.

To quit GM, from the file menu, choose EXIT or in any module screen, choose EXIT. The dialog box as shown as in Figure A-2 appears. Click OK to terminate the program.

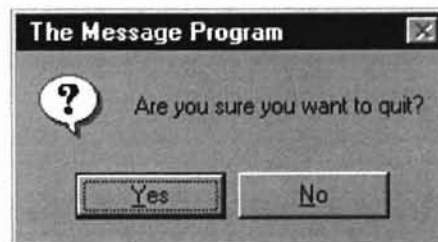


Figure A-2 Exit dialog box.

## Data Entry and Execution of the Program

The main menu window (Figure A.3) consists of three parts. At the top of the window is an “Enter New F<sub>2</sub> Data” area. Below the window is an “Open an Existing File” area. The third part is a group of command buttons.

If you want to enter new F<sub>2</sub> data, you should follow these steps.

1. Choose the “Build” option in the main menu window and click the OK button. The Build window will pop up as in Figure A.4.



Figure A.3 Main Menu Window



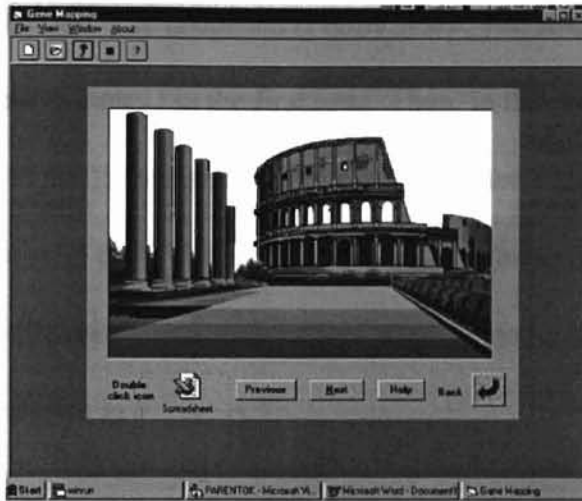


Figure A.4 Build Window

2. Double click on the spreadsheet icon, which is the Visual Basic control for users to open a Microsoft Excel work sheet. The Microsoft Excel page as shown in Figure A.5 will appear.

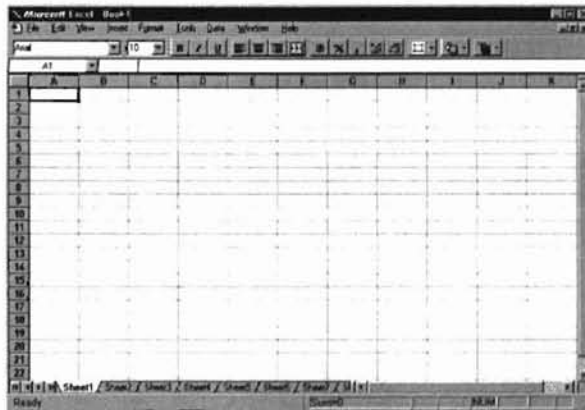


Figure A.5 Microsoft Excel Week Sheet.

3. Click the FILE menu which is located on the menu bar. Select OPEN file, choose A drive, then select " Build 1" file. Double click on it or click it, and then click the OK button. This will open the template shown in Figure A.6.

4. When the “Build1” file is opened (Figure A.6) , you first enter column headings (visible markers ) in the first row. Then, in the cells below, enter data corresponding to pot , plant, and visible marker. Enter x for a dead plant, m for mutant and t for wild type. Use the tab to move the cursor from one cell to another cell. Go to the file menu, and click PRINT to print the input file. An example of an input file is shown in Figure A.7.

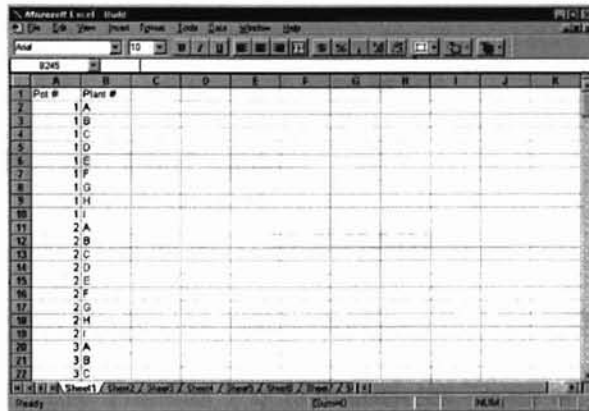


Figure A.6 Build1 File Window.

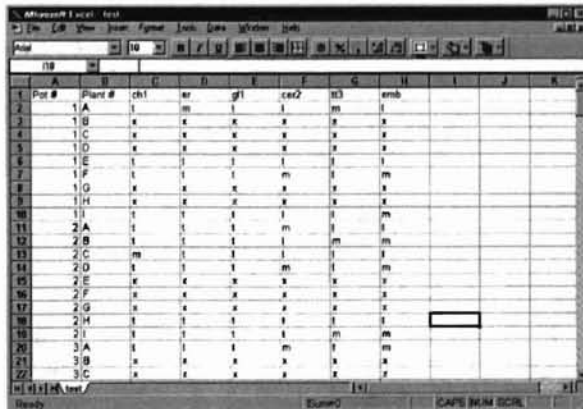


Figure A.7 Example of an Input File

5. Save as file. When you have finished your input data, you should save your file using another name instead of using “Build1” because the “Build1” file is a

read-only template file. If you use the save button or select save in the menu, it takes you to Figure A.8. Use 8 characters for file the name. The file extension name is \*.csv. Click OK to save your input file in A drive.

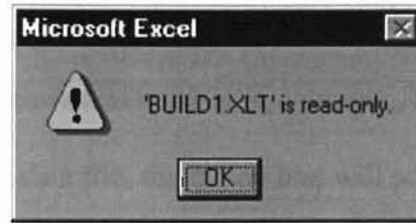


Figure A.8 Save as file.

6. Exit the Microsoft Excel window.

### SUMMARY

There are two ways to run the summary program. One way is to click “Next” on the BUILD window. Another way is to click the “open an existing file” option on the Main Menu, and then click the OK button. The summary will appear as shown in Figure A.9.

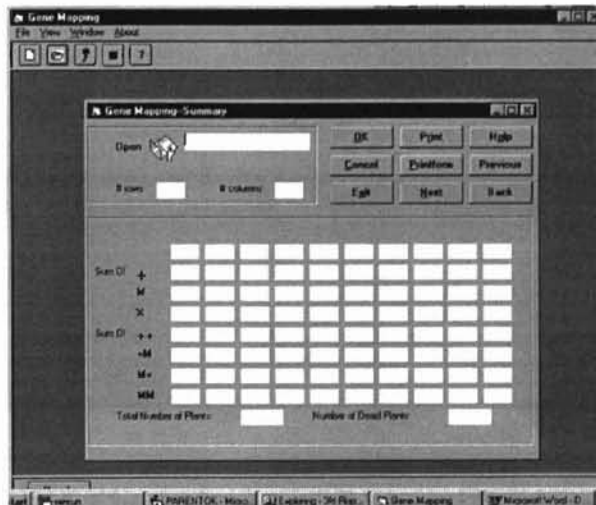


Figure A.9 SUMMARY Window

1. Click the open file icon on the SUMMARY window. The open dialog box will show (Figure A.10). You can choose A drive and open the file just saved in the BUILD module. Click the OK button; the file name will load in the open file box.
2. Enter number of rows and number of columns. If these numbers are different from those in the data file, the dialog box will pop up and let you check the input file again. The number of rows equals the total number of plants plus one. For example, if the number of pots is 27, and there are 9 plants per pot, then  $\#row = 27 * 9 + 1 = 244$ . The number of columns is all visible markers plus 3. For example, if there are 5 visible markers, the  $\#column = 5 + 3 = 8$ .

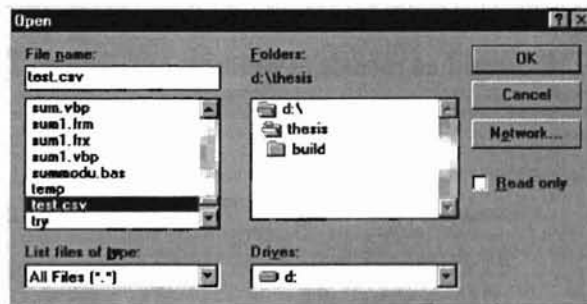


Figure A.10 Summary Open File Dialog Box.

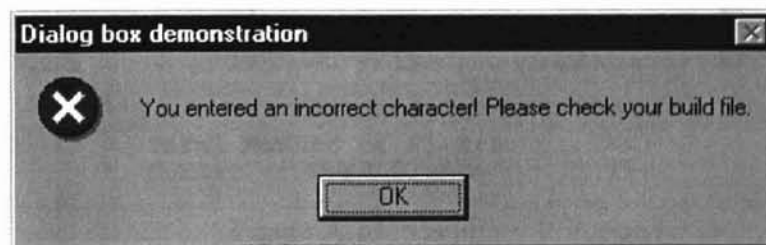


Figure A.11 Check Number of Rows and Number of Columns Box

3. Once the data are entered, the screen should appear as Figure A.12. Press the OK button, and the results are shown on the screen.

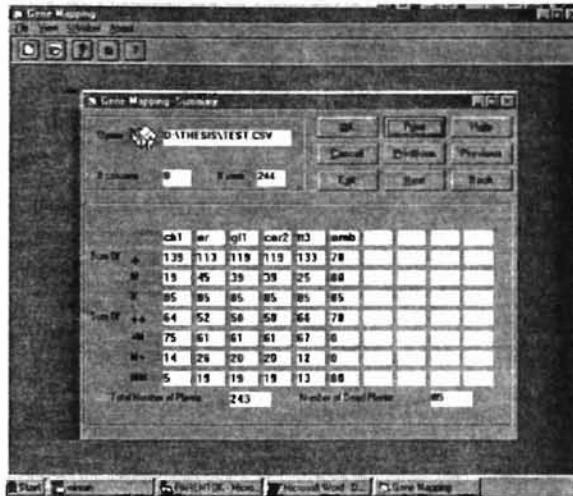


Figure A.12 Summary Results Screen.

4. Press PRINT button, the printout is shown as Figure A.13.

```

Title                A:\test.csv

      ch1          er          g11          cer2          tt3          emb
sum of + 139          113          119          119          133          78
      M 19           45           39           39           25           80
      other 85        85          85          85          85          85

Sum of:
      ++ 64           52           58           58           66           78
      +M 75           61           61           61           67           0
      M+ 14          26           20           20           12           0
      MM 5            19           19           19           13           80

Total Number of Plants: 243
Number of Dead Plants: 85

```

Figure A.13 Summary Printout.

5. Notice that the data file only allows you to enter t, m, x, and b character into data file. If inappropriate characters are used, the dialog box will warn you to check the data file. Click OK, to terminate the process.

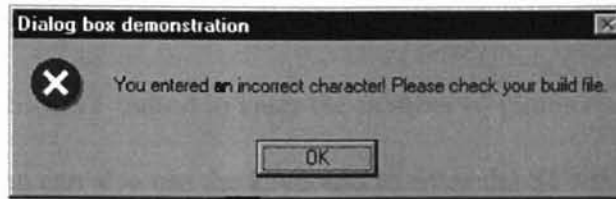


Figure A.14 Check Data File Dialog Box.

6. Once the wrong data are corrected, restart the calculation by repeating step 1.

### CHI – SQUARE

1. In the SUMMARY window, click the NEXT button.
2. To select a title, click on the dropdown list. In option 1, there are 15 visible markers (ch1, tt3, etc.) and 3 multiple marker lines (DP23, DP24, DP28) on this list. Select one of them or simply type in a different marker name in option 2 text box.



Figure A.15 Chi Square Window

1. The A, B, C, D boxes are for the number of plants in each class. You can click DISPLAY button to enter the number of plants corresponding to the title. You can also use the keyboard to enter the SUMMARY results to the Chi-Square window. Once the information is entered, the screen should appear as shown in Figure A.16. To start the calculation, press the OK button.



Figure A.16 Chi-Square Window with Result

4. Click NEXT button to move to the RECF<sub>2</sub> module.

## RECF<sub>2</sub>

1. In the RECF<sub>2</sub> module, click the DISPLAY button; the title and genotypes of the four classes A, B, C, D are shown, as in Figure A.17. You can also enter the title and A, B, C, D information from the keyboard.



Figure A.17 RECF<sub>2</sub>

2. The trial values recombination default is 0.25. You can also enter other data into the trial value box.
3. Click the OK button; the RECF<sub>2</sub> module executes. The results show in the window. See Figure A.18.

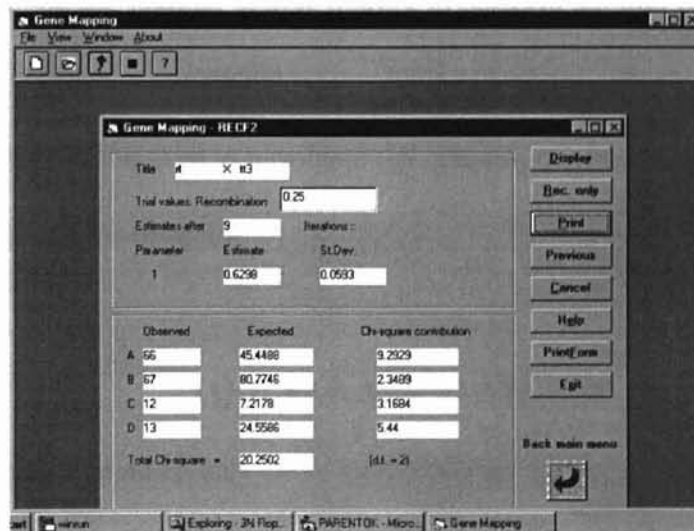


Figure A.18 RECF<sub>2</sub> Window with Results

4. Click the PRINT button; the resulting printout is shown in Figure A.19.



	DP23	X	chl
	Observed	Expected	Difference
	64	39.5	24.5
	75	79	-4
	14	13.16667	0.8333
	5	26.33333	-21.3334
Total	158	158	

CHI\_square = 32.73418 \*\*\*

No.of calculation= 7			
Parameter	Estimate	St.Dev.	
1	0.62633	0.05951045	
	Observed	Expected	Chi-square Contribution
	64	45.313	7.706482
	75	80.68114	0.4000355
	14	7.353664	6.007043
	5	24.65219	15.6663
Total	158	158	29.77986 (d.f.=2)

Figure A.19 RECF<sub>2</sub> Printout.

- Click the PREVIOUS button. It takes you back to the chi-square window.

You can choose another visible marker to perform another calculation until all visible markers are analyzed.

- Click the EXIT button to terminate the process, or choose EXIT from the file menu to exit the program.

## TOOLBAR AND STATUBAR

The toolbar and status bar displayed on the screen are the default options of the program.

- From the VIEW menu, click "hide toolbar;" the toolbar disappears as shown

Figure A.20.

2. From the VIEW menu, click “show toolbar;” the toolbar appears on the screen.
3. From the VIEW menu, click “hide status bar;” then the status bar disappears as shown in Figure A.21.
4. From the VIEW menu, check “show status bar;” the status bar appears on the screen.

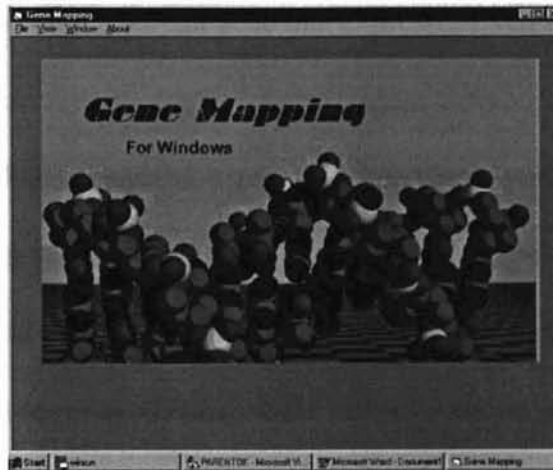


Figure A.20 Hide Toolbar.

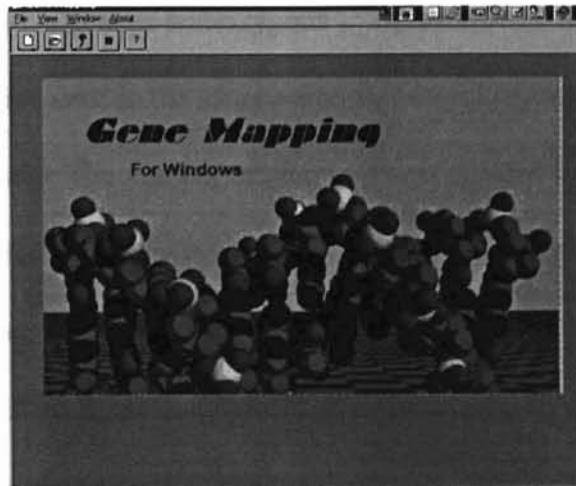


Figure A.21 Hide Statusbar.

## APPENDIX B

### BASIC CONCEPTS OF THE GENE MAPPING[16]

*Genetic Map:* The linear arrangement of mutable sites on a chromosome as deduced from genetic recombination experiments.

*Mutant:* An organism bearing a mutant gene that expresses itself in the phenotype of the organism.

*Recessive gene:* In diploid organisms, a gene which is phenotypically manifest in the homozygous state but is masked in the presence of its dominant allele. Usually the dominant gene produces a functional product, while its recessive allele does not. Therefore the normal phenotype is produced if the dominant allele is present ( in one or two doses per nucleus), and the mutant phenotype appears only in the absence of the normal allele ( i. e., when the recessive gene is homozygous ). By extension, the terms dominant and recessive are used in the same sense for heterokaryons and merozygotes.

*Recombination frequency:* The number of recombinants divided by the total number of progeny. This frequency is used as a guide in assessing the relative distances between loci on a genetic map.

*Backcross:* The cross of an  $F_1$  heterozygote with an individual of genotype identical to one of the two parent individuals.

*$F_1$ :* First filial generation; the offspring resulting from the first experimental crossing

of the plants or animals. The parental generation with which the genetic experiment starts is referred to as  $P_1$  (q.v.).

$F_2$ : the progeny produced by intercrossing or self-fertilization of  $F_1$  individuals.

*Coupling, repulsion configurations* : when both nonallelic mutants are present on one homologue and the other homologous chromosome carries the plus alleles ( a b /++) ( + represents dominant gene) the genes are said to be in the coupling configuration.

The repulsion configuration refers to a situation in which each homologue contains a mutant and a wild-type gene (a+ / +b).

*Locus (plural, loci)*: The position that a gene occupies in a chromosome

*Allele*: one of an array of possible mutational forms of a given gene. When many allelic forms of a gene exist, it is said to show multiple allelism.

*Phenotype*: the observable properties of an organism, produced by the genotype in conjunction with the environment.

*Dominant gene*: see *recessive gene*.

*Recessive lethal*: an allele which kills the cell or organism that is homozygous or homozygous for it.

P value: probability value a decimal fraction showing the number of times an event will occur in a given number of trials.

*Heterozygous*: state of being a heterozygote for one or more gene loci.

*Segregation: the law of segregation*. the factors of a characters are segregated.

In modern terms this law refers to the separation into different gametes

and thence into different offspring of the two members of each pair of alleles possessed by the diploid parental organism.

*Visible Marker:* Mutation that alters the appearance of an organism.

## VITA

Li Wang

Candidate of for the Degree of

Master of Science

Thesis: GUI DESIGN AND IMPLEMENTATION IN GENE MAPPING FOR  
WINDOWS 95

Major Field: Computer Science

### Biographical:

Personal Data: Born in ZhenZho, Henan Province, China. January 10, 1963,  
the daughter of Wenan Wang and Peijun Yang.

Education: Graduated from 19<sup>th</sup> High School of ZhenZho City, Henan  
Province, China in July 1982. Received Bachelor of Science degree in  
Plant Breeding and Genetics Science from Nanjing Agriculture University  
in July, 1986, Nanjing, China. Completed the requirements for the Master  
of Science degree in Computer Science at Oklahoma State University in  
May 1998.

Experience: Graduate teaching assistant from August 1997 - May 1998 in  
Computer Science Department, Oklahoma State University,  
Stillwater, Oklahoma. Research Assistant Professor and Administor  
from July 1986 - December 1994 in Crop Science Institute, Jiangu, Su,  
China.