Development of an Integrated Analysis Tool for the Statistical Genetics Software "Linkage Package"

Keisuke Sugano[1], Mariko Oike[1], Hiroyuki Nishimura[2], and Yukiyasu Iida[2]

(2009年3月7日　受理)

1. Introduction

Recent advances in genetic research are promoting the investigation of polymorphism (i.e., the difference in the genome sequences between individuals) and for applying it to the fields of medical treatment and drug discovery. The process of searching for genes associated with traits (i.e., the presence or absence of genetic diseases, drug effectiveness, and the presence or absence of side effects to drugs) by using genetic polymorphism data of individuals is called trait mapping and it is currently being very actively investigated throughout the world. Statistical genetics analysis methods are effective approach for trait mapping, so statistical genetics is becoming increasingly important.

For linkage analysis in statistical genetics, complicated probability calculations are performed on extremely large amounts of observed data. A computer is used to process the data since would be impossible to process it manually. Dedicated software for statistical genetics analysis is readily available via the Internet. Representative linkage analysis software programs for analyzing disease genes are *Linkage Package, GeneHunter,* and *MAPMAKER/SIBS*[1]. These packages normally run only on UNIX and DOS (i.e., the DOS window in Windows) and they may generate meaningless characters or fail to run properly on the most recent versions of Windows. These software packages were developed in the 1970s and 1980s. They are command-based and have problems associated with the user interface.

In this study, we developed a software program with a graphical user interface (GUI). This program facilitates the creation of two kinds of input files: pedigree files (PedFiles) and linkage parameter files (DataFiles). These files are required input files for the statistical genetic package, *Linkage Package.* We also made it possible to execute everything from the main menu by executing the component programs of the high-speed *FastLink,* which is an extension of *Linkage Package,* on an integrated platform.

Graduate school of Engineering[1], Faculty of Biomedical Engineering[2],
Toin University of Yokohama, 1614 Kuroganecho, Aoba-ku, Yokohama, 225-8502

2. Statistical Genetics Analysis Methods

Statistical genetics refers to all methods of searching for genes associated with specific traits by using statistics. It can be broadly subdivided into linkage analysis and analysis based on linkage disequilibrium (LD). The former is mainly used for pedigree data, whereas the latter is mainly used for group data. Another technique is quantitative trait locus (QTL) analysis. It is a statistical genetics analysis of quantitative trait loci that are associated with such quantitative traits, such as blood pressure, cholesterol level, and height, as well as qualitative traits, such as the presence or absence of genetic diseases, drug effectiveness, and the presence or absence of side effects to drugs.

Linkage analysis can be subdivided into two kinds. One is parametric linkage analysis that requires assuming the mode of inheritance and the other is non-parametric linkage analysis that does not require assuming the probability of transmission from phenotype to genotype (penetrance) . These two kinds of linkage analysis are described below.

2.1 Parametric linkage analysis

Linkage is a phenomenon in which the alleles (allelomorphs) of different loci on the same chromosome are transferred non-independently from a parent to a child. The degree of linkage depends on the recombination fraction between loci. In general, recombination tends to occur more readily the further two different loci are separated from each other. When the recombination fraction between two loci is smaller than 0.5, the two loci are linked. When two loci are not linked, the recombination fraction will be 0.5. Linkage analysis uses this linkage to estimate affected locus for a chromosome. For this estimation, pedigree data is used as observation data with phenotypes of individuals and genotypes at fixed marker loci.

In parametric linkage analysis, affected loci are estimated by assuming the mode of inheritance (autosomal dominant inheritance, autosomal recessive inheritance, etc.), penetrance, and allele frequency. The presence or absence of a linkage is checked by testing a null hypothesis that there is no linkage between affected locus and the marker locus. Actual analysis consists of probability-based estimation and testing. As the function $L\ (\theta)$ of the recombination fraction θ, we first calculate the probability of obtaining the observed data when the affected locus is linked to the marker locus. By considering the ratio of $L\ (\theta)$ to the probability ($L\ (\theta = 0.5)$) at free recombination, we calculate the value of θ at which the probability ratio is a maximum. In addition, a test is performed based on the probability ratio. The common logarithm of this probability ratio is called the LOD score. A linkage is usually determined as being present when the LOD score is 3.0 or greater. The affected locus is estimated from the recombination fraction.

Software programs for performing parametric linkage analysis are *Linkage Package*, *FastLink*, and *GeneHunter Linkage Package* was developed in Pascal[2]. *FastLink*, which uses an improved algorithm from *Linkage Package* to give a faster processing speed, is written in C. *GeneHunter* conducts analysis by using a mathematical method called the hidden Markov model. This analysis

program can conduct rapid multipoint analysis on many marker loci, but its application is limited to small pedigrees.

These software programs are available from the website of Rockefeller University (http://LINKAGE.rockefeller.edu/soft/).

2.2 Non-parametric linkage analysis

For non-parametric linkage analysis, the inheritance mode or penetrance is not assumed. Instead the number of homologous alleles shared between affected members of a pedigree is used. Causal disease genes are identified by the affected sib-pair method. Identical by descent (IBD) refers to the sharing of alleles from a common ancestor between two members of a pedigree. Non-parametric linkage analysis involves searching for loci that deviate greatly from the expected distribution obtained by assuming that the distribution of the number of IBD alleles has no linkage with affected loci. *MAPMAKER/SIBS* is an analytical software program based on this method. This program is also available from the website of Rockefeller University (http://LINKAGE.rockefeller.edu/soft/).

3. Outline of *Linkage Package*

3.1 Configuration of *Linkage Package*

Linkage Package is typical software for parametric linkage analysis. For genetic analysis using this software, a pedigree file (PedFile) that contains data about family members (such as parent-child and siblings) and infected members in the family, and a linkage parameter file (DataFile) about allele frequency and recombination rate are set and input into the genetic analysis software. Figure 1 shows the configuration of *Linkage Package*. The support programs for creating input data are *Makeped and Preplink*. These programs are used to create pedigree data and linkage parameter data.

The main analysis programs are *Unknown, ILink, MLink,* and *LinkMap*. *Unknown* checks data conflicts. *MLink* and *ILink* are programs for single-point analysis. *MLink* varies the recombination fraction iteratively within a fixed range, and calculates the probability and the LOD score. *ILink* calculates the

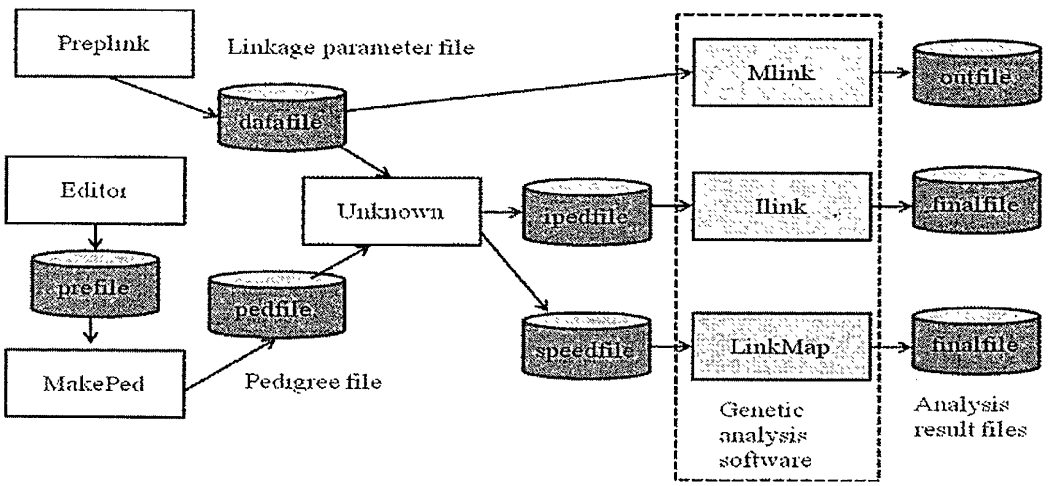

Figure 1. The configuration of Linkage Package

probability and the LOD score by the optimum estimation method. *LinkMap* is a multipoint analysis program and calculates the LOD score by moving a affected locus among several marker loci.

3.2 Input File Format

Makeped is used to create a pedigree file from data in text format created using a text editor. Information about an individual is written on one line. For this information, the following parameters are written sequentially (space delimited) :

1) Pedigree ID
2) Individual ID
3) Individual's father ID (0 if no father in the family)
4) Individual's mother ID (0 if no mother in the family)
5) Sex (Male: 1; Female: 2)
6) Phenotype
7) Allele number

The above pedigree information and the allele numbers of object loci are written. When the phenotype of 6) indicates a diseased state, the value is 2 for "*Affected*", 1 for "*Not affected*", and 0 for "*Unknown.*" Figure 2 shows an example of a pedigree (the part of records in prefile in Fig.1).

A linkage parameter file is generally created by using Preplink and can also be created directly by using a text editor. The following information is written in a linkage parameter file:

1) General information about loci and their orders (number of loci, mutation rate, order on chromosome, etc.)
2) Locus type (quantitative variable, affection status, binary factor, and allele numbers)
3) Information about recombination (recombination rate, difference by sex, and interference by map function)
4) Required information for program (depends on the specific program)

4. Development of Integrated Analysis Tools Using GUI

Using Visual *C#* [3], which enables GUI applications to be rapidly developed, in this study we developed a user interface to easily create two kinds of input files: pedigree files (PedFile) and linkage parameter files (DataFile). These files are necessary input files for most genetic statistical software. For the high-speed statistical genetics software *FastLink* (which was extended from *Linkage Package*) *MLink, ILink,* and *LinkMap* were revised to be run using Visual *C++*[4], which is grammatically compatible with Visual *C#*. In addition, all of these software programs were made to be executable from the main

Pedigree ID	Person ID	Father ID	Mother ID	Sex	Phenotype	Allele Numbers	· · · ·
1	1	0	0	1	2	1 2	
1	2	0	0	2	1	1 1	
1	3	1	2	1	1	1 1	
1	4	1	2	2	1	1 1	
1	5	1	2	1	2	1 2	

Figure 2. An example of pedigree data

menu (Fig. 3). The configuration of the integrated environment system based on the integrated platform allows the above-mentioned linkage programs, which were created in Visual C++, to be activated from the menu by specifying their execution files to start up *Unknown, MLink, ILink,* and *LinkMap* (Fig. 4).

4.1 Development of pedigree data creation program

For pedigree data input, a pedigree chart is created by a GUI operation from components representing affected and non-affected males and females. By clicking a graphic representing an individual (○ or □), the allele setting of the selected person is displayed on the screen (see Fig. 5) . The pedigree data input file creation program checks visual pedigree chart information for conflicts in the pedigree and creates a file of the PedFile format in the end.

4.2 Development of linkage parameter creation program

Preplink is a program of about 2,000

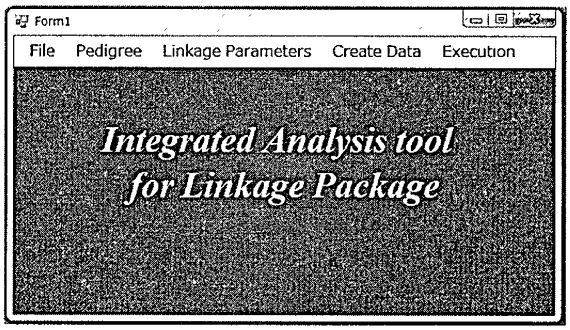

Figure 3. Main menu of Integrated Analysis tool

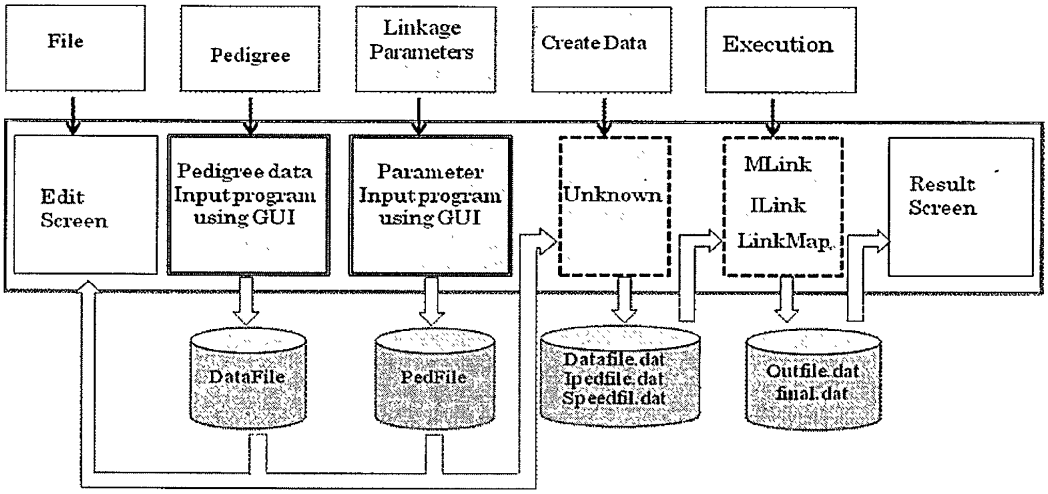

Figure 4. The configuration of Integrated Analysis tool

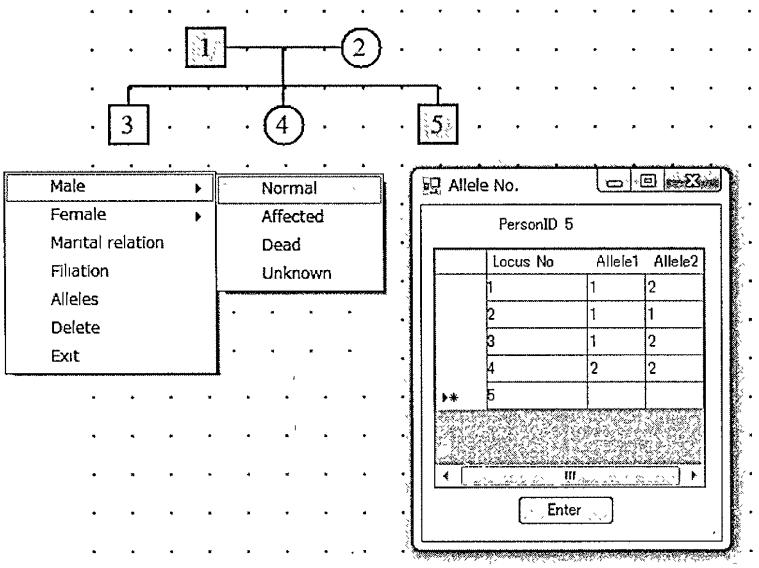

Figure 5. The input of pedigree data using GUI

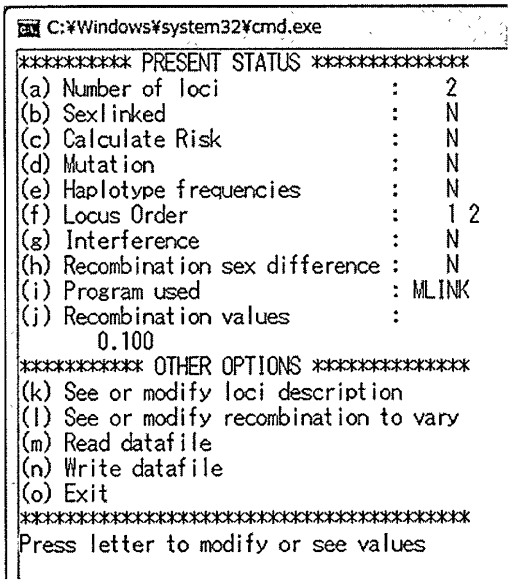

Figure 6. The input of linkage parameters in PrepLink

lines written in Pascal. For porting to *C#*, we compiled the program using a program developed at our laboratory that automatically translates programs from Pascal to *C*. The C code was verified to run in Visual *C++* and it was then ported to *C#*.

The original *Preplink*, which was written in Pascal, was enabled a user to select the symbols of input items from a menu and enter them sequentially in a command

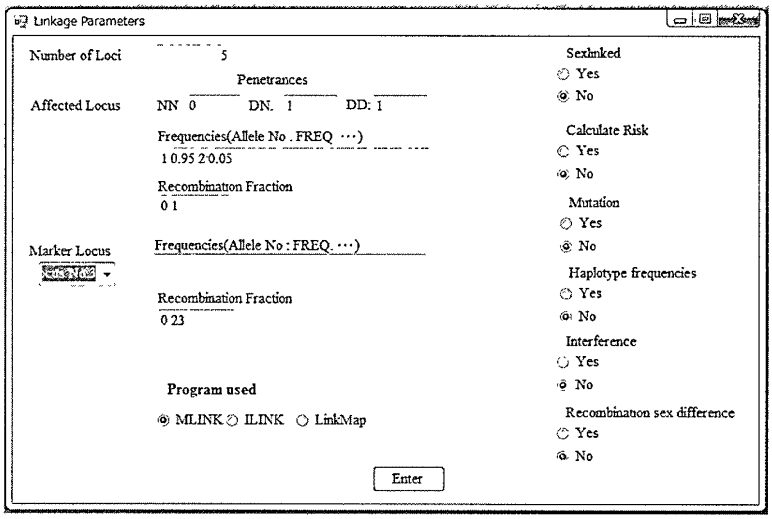

Figure 7. The input of linkage parameters using GUI

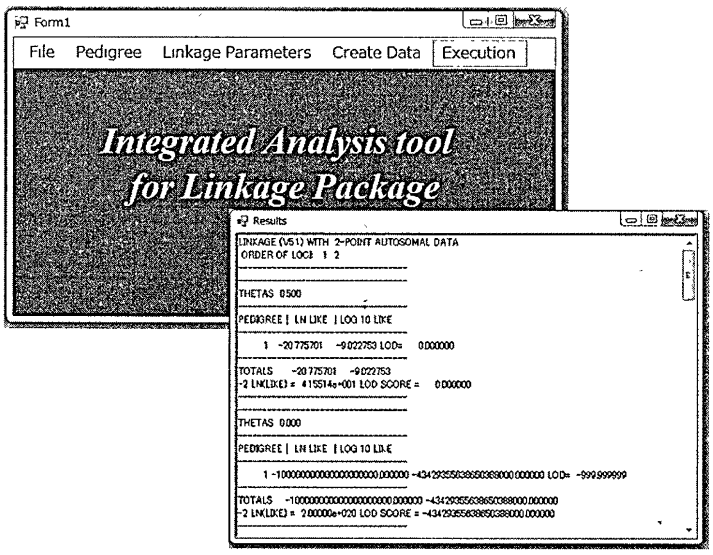

Figure 8. The execution of linkage analysis clicking the main menu button

screen. For example, to input new allele data for gene loci, complicated input screens were needed from which "(a)Number of loci" was selected and then the number of alleles and frequency were entered (Fig. 6).

The linkage parameter creation program with GUI allows all settings to be displayed on a single screen and provides listboxes for gene loci and radio buttons for "Yes" or "No" to simplify input (Fig. 7).

4.3 Execution of linkage analysis

For data files created by the programs described in Sections 4.1 and 4.2, *Unknown* is read and executed by clicking the analysis data creation button in the main menu to create the final analysis data.

The linkage analysis button is then clicked to execute *MLink, ILink,* or *LinkMap* specified by parameter settings and to display its results. By clicking the main menu button, linkage analysis can be executed and the results would be shown in a screen efficiently (Fig. 8).

5. Conclusion

The data processing programs, *Makeped* and *Preplink* can be operated by GUI in the latest object-oriented language *C#*. *Unknown, ILink, MLink,* and *LinkMap* play a central role in analysis. Therefore, their *FastLink* versions were ported from ANSI *C* to Visual *C++* to enable them to run optimally on a Windows PC.

The data creation program for analysis is about 1,100 lines in *C#*. When *Preplink* was created in *C#*, the actually created code was about 1,200 lines long. The analysis tool developed to efficiently read these programs and the main programs (such as *Unknown* and *MLink*) was about 3,000 lines long. About 40% of the program lines in the code was automatically generated by the rapid application development function of Visual Studio *C#*.

The input efficiency by GUI is currently being evaluated. The *C++* execution files of *MLink, ILink,* and *LinkMap* are currently read and executed as processes. In the future, we intend to port their codes to *C#* for high-speed processing and also work with *GeneHunter and MAPMAKER/SIBS,* in addition to *Linkage Package.*

References

1) Jurg Ott , Analysis of Human Genetic Linkage third edition.Johns Hopkins press, 1999
2) Niklaus Wirth,Algorithms + Data Structures = Programs. Prentice-Hall, 1975,
3) Anders Hejlsberg、Mads Torgersen、Scott Wiltamuth、Peter Golde,The C# Programming Language ,Microsoft .Net Development Series,2006.
4) David Kruglinski、Scot Wingo and George Shepherd,Programming Microsoft Visual C++,Microsoft Press,1988