[#]Table of Contents

1. Overview

1.1 Methodological Summary

1.2 Choice of Tasks, Algorithms and Parameters

1.3 Importing Data

2. General Application Structure

2.1 Pedigree Reconstruction Tools window

Output Memo
Output Grid

2.2 Control Panel window

Single Generation Tasks tab sheet
Histogram Display tab sheet
Bootstrap Coverage Display tab sheet
Genotype Chart tab sheet
Allele Chart tab sheet

2.3 Memory Management panel

2.4 Run Script button

3. Defining Application Tasks

3.1 Single Generation Tasks tab sheet

Task Flowchart

3.1.1 Single Generation Tasks tab sheet (Algorithm Details)

Task panel
Select Enumeration Algorithm panel
MSG Algorithm Options panel
Triplet Enumeration Options panel
Partition Options panel

[#] Table_of_Contents

Genotype Error Iteration panel
Bootstrap Options panel
Randomization panel
Output Options panel

3.1.2 Single Generation Tasks tab sheet (Simulation Details)

Simulation Type panel
Sample from Individuals panel
Frequency Properties panel
Sibling Group Sizes panel

3.2 Histogram Display tab sheet

3.3 Bootstrap Coverage Display tab sheet

3.4 Genotype Chart tab sheet

3.5 Allele Chart tab sheet

[#]Section 1. Overview

Version 2.2x of PRT represents a new implementation of algorithms for sibling reconstruction based on algorithms described in the following publications:

- Almudevar, A, Field, C. (1999). Estimation of single generation sibling relationships based on DNA markers. Journal of Agricultural, Biological and Environmental statistics, **4**, 136-165.
- Almudevar, A, (2001). A bootstrap assessment of variability in pedigree reconstruction based on DNA markers. Biometrics, **57**, 757-763.

It is assumed that genotypic data in *L* common loci is available for *N* individuals from a single generation. The problem is to partition the individuals into sibling groups (SG), often referred to as Sibling Reconstruction (SR) (as opposed to parentage assignment). The problem is statistical in nature in that perfect resolution into the true SGs is generally not possible, but high accuracy is possible with data of sufficient quality.

The underlying methodological concepts are discussed in **1.1 Methodological Summary**. In **1.2 Choice of Tasks, Algorithms and Parameters** the available tasks and options are introduced, while in **1.3 Importing Data** file importing and formatting are discussed. The remaining sections provide more specific detail in a reference format.

Next Subsection
Return to Table of Contents

[#] **Section_1._Overview**

[#]1.1 Methodological Summary

The methods implemented here rely primarily on genetic exclusion. A subset of individuals is called a Feasible Sibling Group (FSG) if they are genetically compatible as siblings based on the available data. Of course, they may or may not be true siblings. A FSG is called Maximal Sibling Group (MSG) if it is not a strict subset of another FSG. Thus, a list of all MSGs forms the most compact representation of all FSGs, since any FSG must be a subset of at least one MSG. It is important to note that two individuals always form an FSG, so MSG enumeration may be confined to subsets of at least 3 in size.

Surprisingly, construction of an MSG list is feasible using data of a quality typically available. Two MSG enumeration algorithms are available in PRT, termed MSG-G (graph based) and MSG-T (triplet enumeration). They give the same output, but using very different enumeration strategies. Both are described in detail in Almudevar and Field (1999), but we highlight the essential distinction here. MSG-G may be thought of as a 'top down' algorithm in that it starts from larger subsets and systematically splits them into MSGs. On the other hand, the MSG-T algorithm is a 'bottom up' algorithm, since it starts with all FSGs of size 3, and systematically consolidates them into larger FSGs, eventually yielding all MSGs. As might be expected, the MSG-G algorithm is more efficient when larger SGs are present, but the MSG-T is more efficient when the sample is characterized by smaller SG sizes or sparse relatedness in general.

A modification of the MSG-G algorithm was also proposed in Almudevar and Field (1999) for use when the computational burden is too great, which is denoted MSG-MG (modified graph based). One useful property of the MSG-G algorithm is its ability to efficiently restrict MSG enumeration to MSGs of a size no smaller than a given threshold *T*. The MSG-MG proceeds by first enumerating MSGs using threshold *T*, then constructing a partial SG partition from the output. The threshold is then lowered, and the algorithm applied again to the data, but with any previously estimated SGs removed. This process continues iteratively until the partition is complete.

Although the recommended choice of algorithm depends also on the quality of the data, the following rules can serve as a rough guide:

Table 1 When to use MSG-G, MSG-T or MSG-MG.

Use:	When:
MSG-T (triplet enumeration)	The largest SG size is ≤ 5 **OR** there is only 1 SG of size ≤ 15 within many unrelated individuals.
MSG-G (graph based)	The largest SG size is 5 - 25.
MSG-MG (large SG modification)	The largest SG size is > 25.

[#] **1.1_Methodological_Summary**

Enumeration of an MSG list is only an intermediate step, since an estimate must be a partition of the individuals into separate FSGs. The list of MSGs will usually contain overlapping sets, so a partitioning algorithm is applied by PRT automatically to the MSG list. This is based on a heuristic algorithm that can assign a score to any given FSG. The partition algorithm first searches all MSGs for subsets which may statistically resemble true SGs more than the MSGs. At or below a given threshold all subsets are searched, and above only subsets obtainable by deleting 1 or 2 individuals are considered (this number is optional). The algorithm then assembles a partition from these subsets. These parameters are specified in the **Partition Options** panel (see section **3.1.1**). Although this score is based on the likelihood function, it is differs in some important ways, and the resulting partition is quite different from a maximum likelihood estimate. The reader is referred to Almudevar and Field (1999) for more detail.

Next Subsection
Return to Table of Contents

[#]1.2 Choice of Tasks, Algorithms and Parameters

In this section we describe how to define a job. This is done mostly in the **Single Generation Tasks** tab. There are 12 separate panels of options, and not all are needed for a specific job. It will simplify things considerably to think of a job as being defined by the topmost three panels:

- **Task** panel
- **Select Enumeration Algorithm** panel
- **Simulation Type** panel

To make the application a bit more user-friendly, whenever any of these three panels is changed the title label of panels containing parameters or options relevant to that job will be left in bold font, while others will be grayed out (the panels can still be edited). For the **Output Options** panel this prompting extends to individual options (see section **3.1.1**). The panel precedence is given as a flow chart in section **3.1**

Tasks **(1)** and **(2)** (estimation and bootstrap) are performed on an input data set. There are default settings for the required parameters which would be sensible choices for most tasks, but they are worth getting to know, eventually. Usually, they control a tradeoff between accuracy and computation time. Tasks **(3)** and **(4)** are simulation studies. These allow the user to define a sibling and genotype data model to assess the accuracy and computation time of a selected algorithm. The data can be simulated from a probabilistic model, or based on data provided by the user (either allele frequencies or actual genotype samples). Determining the number of loci needed for acceptable kinship resolution is more difficult for SR than for parentage assignment, so this feature is intended to be used in the design stage of a study as well as for the subsequent data analysis. We next give brief highlights of the four tasks, more detail can be found in the rest of the manual.

(1) Estimate Partition. This is the primary task. The user first imports data from a file using the **Genotype Chart** (the user may also enter the data manually, if so inclined). The enumeration algorithm is then selected in **the Select Enumeration Algorithm** panel. Guidelines for this choice are given in Table 1 (in the absence of a recommendation, the safest choice would be the MSG-MG algorithm, with the **Use Suggested Settings** check box selected). PRT 2.x provides a comprehensive selection of reporting options in the **Output Options** panel. See section **3.1.1** for more detail.

(2) Bootstrap Partition. This is used like task **(1)** but performs a bootstrap estimate of the algorithms' accuracy (the algorithm should therefore be the same). There is no estimate summary or reporting options, so this task would naturally be used after performing task **(1)**. The output consists of an estimated sampling distribution of the error, expressed as a partition distance between the true and estimated partition. The distribution is output numerically in the **Output Memo** and displayed in the

[#] **1.2_Choice of Tasks,_Algorithms_and_Parameters**

Histogram Display tab. Note that the true partition is not (and need not be) known, so that the sampling distribution is itself a statistical estimate, the accuracy of which can be assessed using task **(4)**.

(3) Evaluate Partition with Simulations. The algorithm to be assessed is selected as in task (1). The user then selects a simulation model from some combination of the **Simulation Type**, **Sample From Individuals**, **Frequency Properties** and **Sibling Group Sizes.** More detail is given in section **3.1.1**, but we make a crucial distinction here. When the **Single Trial** option is checked one data set is simulated, then treated the same way as in task **(1)**, using all the available reporting options. Furthermore, the simulated data remains in the **Genotype Chart** and may be saved, edited or otherwise used again. If the **Single Trial** option is not checked, then the model will be simulated *N* times (which is the number entered into the **N Repeated Trials** text box). The output will consists of the distribution of the partition distances between the estimates and the true partition. This is given numerically in the **Output Memo** and displayed in the **Histogram Display** tab.

(4) Evaluate Bootstrap with Simulations. This will probably be the least familiar option. This task assesses the accuracy of the bootstrap procedure of task **(2)**, which is an estimate of the sampling distribution of the partition estimation error. Since this is itself a statistical estimate, it may or may not be accurate. The algorithm follows the procedure described in Almudevar (2001). The output is a plot of the nominal quantiles of the error distribution against the actual values (available by sampling from known simulated models). Ideally, this graph should be the identity. When it's not, it's usually above the identity. In this case the procedure is conservative, which is acceptable if not ideal. The procedure should not be used when the graph is below the identity. Fortunately, in the authors' experience this doesn't happen very often, but more insight would always be welcome. The graph is displayed in the **Bootstrap Coverage Display** tab, and the (X,Y) data can be output to the **Output Memo** by selecting the **Output Bootstrap Plot XY Data** option in the **Output Options** panel. See section **3.1.1** for more detail. Note that this list can be quite long. Also note that the procedure itself is, necessarily, very time consuming.

Miscellaneous comments on available options and parameters panels follow next. See sections **3.1.1** and **3.1.2** for more detail.

Bootstrap Options. The bootstrap procedure resamples putative parental genotypes, and there are several ways to do this. Experience suggests that the **Condition on Alleles** option works best, but a theory-based resolution of this issue is still pending. More than one option can be used at a time, and the graph will superimpose all methods selected. Of course, increasing the **Number of Replications** parameter gives a more accurate procedure, but a smaller number can be used initially to estimate computation time.

Partition Options. All enumeration algorithms are followed by the partition algorithm, and the total computation time is very sensitive to the choice of parameters. With SGs of a size in the hundreds, this is especially true of the **Reduced Search Deletion** parameter. This can be set in the range 0-2,

but it is not recommended to set it to 0. This is because it is possible that unrelated individuals spuriously form FSGs with true SGs, and this might be detectable with the type of statistical comparison used in the partition algorithm. An example of this can be found in the Quick Start Task 1 example provided with PRT 2.x.

Randomization. All tasks except **(1)** use a random number generator. If the **Specify Random Seed** option is checked, then the application will use the random seed specified by the user, otherwise a clock generated number is used.

Genotype Error Iteration. PRT is able to flag potential genotype errors. It does this by systematically examining each locus to determine whether or not larger sibling groups can be formed after deleting (by reclassifying as missing) some bounded number of genotypes. If so, this is interpreted as evidence of a genotype error. PRT will optionally report flagged genotype errors following a partition estimate (select this in the **Output Options** panel). In addition, the error flag can be incorporated into the partition using an iterative modification of the algorithm. Flagged genotypes are deleted from the data following each iteration, then the partition is estimated again with the modified data. See the **Genotype Error Iteration** panel for more details.

MSG Algorithm Options. This panel is used only for the MSG-MG algorithm. If the **Use Suggested Settings** option is checked then the parameters are selected by the program (see below for details). So, why would you want to set your own? (Uncheck the **Use Suggested Settings** option to enter this territory).

(1) The algorithm can be used simply to detect the presence of large sibling groups. If this is what you want to do, then set **Minimum MSG Size** to the minimum size you wish to detect, then set the Number of Iterations to 1 (the remaining parameter is not used when only 1 iteration is specified). If you do this, you may want to check the **Output MSG List** option, since this gives the output of the MSG enumeration algorithm directly in the **Output Grid**.

(2) If the memory is overwhelmed even with the suggested settings, then try setting **Minimum MSG Size** to the total number of individuals, and decreasing **Min MSG Size Increment** (the suggested setting is 10). If this doesn't work, it's always possible that the SGs are actually too small, so that MSG-T, being designed for this scenario, might work better. Otherwise, the remaining suggestion would be to split the sample into subsets, and partition each separately (the **Genotype Chart** can be used to restrict the input in this way). The **Output Nonexcluded Parents** option will report the putative parents, permitting reassembly of the partition, with a bit of patience.

(4) The computation time will decrease somewhat if **Minimum MSG Size** is set to a number just below the largest true SG size (assuming this can be estimated).

(5) The MSG-MG is a heuristic procedure, and we conjecture that it is less accurate the more

heuristic it is, although a formal proof is not available. To make the procedure less heuristic decrease **Minimum MSG Size** and/or increase **Min MSG Size Increment**. Also, make sure **the Number of MSG Iterations** is large enough to reach a threshold of 3.

Triplet Enumeration Options. This panel is used only for the MSG-T algorithm. This algorithm is pretty straightforward, and works in only one way. The main issue here is whether the MSG-T or the MSG-G (or MSG-MG) is the appropriate algorithm. If you wish to determine the largest MSG in the sample, and suspect that it might be small, then uncheck the Unlimited Iterations option, and set **Maximum Triplet Iterations** to, say, 1. The program will either report the size of the largest MSG (in which case the partition algorithm has actually completed, and there's nothing more to do), or it will report that the largest MSG must be at least some number in size. If that number is large (say 6 or more), or the task has taken a long time, you might wish to consider the MSG-G algorithm. Otherwise, you can increase **Maximum Triplet Iterations** and try again.

Output Options. A number of report options are available. The following:

Output Genotype Error Screen
Output Alternative SG Assignments
Output Nonexcluded HSGs
Output Nonexcluded Parents

are output to the **Output Memo** as text. This is true also of **Output Bootstrap Plot XY Data**. This gives the (X,Y) data used to plot the graph in the **Bootstrap Coverage Display** tab when using task **(4)**. It can be lengthy.

Output MSG List is selected to output the enumerated MSG list to the **Output Grid**. This is normally not needed for an estimate, but it does form the basis of the whole procedure, and so is worth scrutinizing now and then as a good way to gain insight into the procedure.

PRT normally reports estimated SGs using row indices associated with the **Genotype Chart**. If individual labels are present, and you'd rather use those, select the **Use Labels for Partition Summary** option. Make sure that the label column is specified in the correct field on the **Genotype Chart** page.

Simulation Type. The **Single Trial** option is explained above. It is a nice way to become familiar with PRT without needing an omnibus collection of data sets. Remember that with this option the simulated data remains in the **Genotype Chart**, and can be used as though it had been imported.

There are 4 ways to simulate data. More detail is given below. Option (1) is fundamentally different from the others in that it relies on user provided genotypes from a sample in which the true SGs are known. The SG labels must be included and are identified to the program using the correct field on the **Genotype Chart** page. This technique was used Almudevar and Field (1999) to generate test data

sets. A simulated sample consists of a random sample of rows from this data set. The size of this sample is specified in the **Sample From Individuals** panel.

The other three methods use 'gene-dropping', and differ only in the way population allele frequencies are defined. A pedigree is defined in the **Sibling Group Sizes** panel, genotypes are simulated for the parents according to population genotype distributions (HWE, of course), then simulated for the offspring using Mendelian laws.

The user may import allele frequencies for use in simulations, using Option (2). This is done using the **Allele Chart**. Otherwise, Options (3) or (4) specify a uniform or Zipf allele frequency distribution. The number of alleles per locus and the number of loci must then be specified in the Frequency Properties panel.

A quick word about allele frequency distributions. The use of the uniform distribution in simulation studies is quite widespread (the authors are no exception). This might be useful in standardizing simulation studies, but it isn't very realistic. The Zipf distribution is proportional to $1/i$, $i = 1, \ldots, n$ for n alleles. Our own experience is that this comes pretty close to actual allele frequencies in many cases, and requires no more information to define completely than does the uniform (that is, the number of alleles). So both are included.

Sample From Individuals. See **Simulation Type** panel.

Frequency Properties. See **Simulation Type** panel.

Sibling Group Sizes. The pedigree structure of a simulation model is defined here. Each row represents a "half sibling group complex", which is repeated *N* times. Use only columns *N* and *Gr1* if there are no half siblings. Otherwise, see section **3.1.1** for more detail.

Next Subsection
Return to Table of Contents

[#]1.3 Importing Data

Data sets are mostly managed in the **Genotype Chart** tab. When using imported allele frequencies the **Allele Chart** tab is used, which has a similar but simpler structure, so the following preamble applies here as well. More detail is given below, but we will introduce the basics here.

The data is assumed to have a straightforward genotype matrix structure. Rows represent individuals, and genotypes are represented in allele pairs consisting of consecutive columns representing *L* loci. Formatting is fairly flexible. Any number of columns may be included, as long as the genotype data appears as 2*L* consecutive columns, and in consecutive rows, and alleles are represented as nonnegative integers. Delimiters may be mixed, and are actually taken to be any character other than a letter, a digit or one of these [_;:\] , not including the square brackets (that is, a square bracket is a delimiter). Consecutive blanks are interpreted as a single delimiter.

Once the data is entered, the location of the genotype data, and any labels (individual or family) should be specified in the **Genotype Chart**. Header rows are not a problem, as long as they are properly delimited. Once the row and column definitions are entered, press the Relabel button, to make sure you and the program are in agreement as to the location of the data. It is possible to specify a consecutive subset of the rows, in which case any task will use data only from those rows (press Relabel to make sure you have the right rows). Note that the row indices appear on the left of the grid will be those used to report the partition, unless the **Use Labels for Partition Summary** option in the **Output Options** panel is selected.

When data is imported, or when simulated data is written to the **Genotype Chart**, space is made for a column into which an estimated partition will be placed in labeled form. This is indicated in the **Estimate SG labels in column:** text box, which can be changed. If set to zero, this feature is not used.

Return to Table of Contents

[#] **1.3_Importing_Data**

[#]Section 2. General Application Structure

The application is structured as two windows. The primary **Pedigree Reconstruction Tools** window handles basic application functions. It also contains two child windows **Output Grid** and **Output Memo** for various forms of output. Application jobs are defined and initiated from the **Control Panel** window. Data is also input and displayed from this window. Having separate windows for input and output seems to work well for GUI based applications (although specialized graphical displays remain in the **Control Panel**).

An overview of the window system follows, with further detail given below.

Return to Table of Contents

[#] **Section_2._General_Application_Structure**

[#]**2.1 Pedigree Reconstruction Tools** window

This window has a menu system of basic system functions, as well as two child windows:

Output Memo. General text output for all application jobs appears here.
Output Grid. Structured output for various applications jobs appears here.

The **File** menu can be used to read data into the **Genotype Chart,** although this can also be done directly from **Genotype Chart**, which is probably more convenient. Contents of the **Output Memo**, **Output Grid** and **Genotype Chart** can also be saved from this menu. The **Edit** and **View** menus contain standard controls, some also available by right-clicking the mouse. The **Output Memo** can be manually edited, but the **Output Grid** cannot be. The **Control Panel** window can be hidden and unhidden using the **View** menu. The **Help** menu gives access to the help file (what you are reading now), as well as the PRT version number.

The **Output Grid** is used for the output of the enumerated MSG list. This feature is enabled by selecting the **Output MSG List** option in the **Output Options** panel in the **Control Panel** window (**Single Generation Tasks** tab).

The **Output Memo** will be used quite frequently for task reporting. Input and task information is given, as well as most requested output, including the pedigree itself (which will also be output to the **Genotype Chart**). An example is given below. The report includes a start and finish time and date, and also the computation time of the algorithm itself in milliseconds.

```
============
> Started at 3:02:48 PM on 6/14/2011
> Task
>    (1) Estimate Partition
> using:
>    C:\USER\msg_2010\table.falcon
>    Genotypes in rows:1 to 9
>    Genotype in columns:1 to 18
>    Indiv labels in columns:0 to 0
>    Group labels in columns:0 to 0
>    Estimated SG labels in column:19
> Algorithm: (3) MSG-MG (large SG modification)
> MSG Algorithm Options
>    Use Suggested Settings
>    Maximum Partition List Size=3000
>    Max Sib Size for Full Search=10
>    Reduced Search Deletion=1
>    No missing value
Estimated Partition =
```

[#] **2.1_Pedigree_Reconstruction_Tools_**window

```
(1) 3 6 7 8
(2) 1 2 4
(3) 5 9
>end
> Completed at 3:02:48 PM on 6/14/2011
> Elapsed Time (ms): 94
```

Return to Table of Contents

[#]**2.2 Control Panel** window

The upper portion of the Control Panel displays six separate tab sheets, while the lower portion (from the **Go** button downwards) remains visible and functional for each tab sheet:

Single Generation Tasks. Define the specific job and parameters in this tab sheet.
Histogram Display. Various tasks can be summarized by a discrete probability which is displayed in histogram form in this tab sheet.
Bootstrap Coverage Display. The bootstrap evaluation task constructs an estimate of the functional relationship between nominal and actual coverage of a confidence set, which is displayed in this tab sheet.
Genotype Chart. Genotype data is displayed in a spreadsheet gird in this tab sheet. Data may be read from a file or entered manually. In addition, data may be simulated, stored in this spreadsheet then saved in a file.
Allele Chart. When allele frequencies are needed, they are read into a spreadsheet in this tab sheet.

Go button

All jobs are initiated by pressing the Go button, which is visible and active from all tab sheets of the **Control Panel** window.

Cancel button

This button interrupts a job, as politely as possible. Partial results are reported, where feasible.

Hide Window button

This button hides the **Control Panel** window, which may be unhidden from **View** menu of **the Pedigree Reconstruction Tools** window.

Show Main Form button

Brings the Main Form (**Pedigree Reconstruction Tools** window) into view.

Return to Table of Contents

[#] **2.2_Control_Panel_**window

[#]2.3 Memory Management panel

The various forms of MSG algorithms are essentially constructive enumeration algorithms, and so potentially require large amounts of memory. While older versions of PRT allocate memory dynamically, version 2.2x will allocate a fixed amount of memory once when the program is invoked, to be used by the MSG-G, MSG-MG and partitioning algorithm, while MSG-T continues to allocate memory dynamically. There are three memory allocation parameters, which have default values that may be changed by the user. These are (with default values):

tree_memory 1000
max_msg 3000
max_partition 6000

The value **tree_memory** specifies the memory in megabytes (MB) allocated to the construction of the MSG tree (750 MB is sufficient to apply the MSG-G algorithm to the complete Atlantic Salmon data set originally reported in Almudevar and Field (1999)). The other parameters indicate the maximum partition list size and MSG size. The value entered as **Maximum Partition List Size** in the **Partitions Options** panel cannot exceed **max_partition**, and **max_msg** should generally be at least as large as the number of individuals in a sample. The program will terminate without completion if a MSG larger than **max_msg** is found. A total of 2***max_partition***max_msg** bytes is allocated to the partitioning algorithm.

To change the memory allocation parameters, create a text file called **msg_proj_memory_settings.txt** in the same directory as the application, containing the lines shown in this section, with the desired settings replacing the default values shown. The PRT 2.2x distribution comes with a file **msg_proj_memory_settings_example.txt** which may be modified for this purpose.

[#] 2.3_Memory_Management_Panel

[#]**2.4 Run Script** button

PRT 2.2x now supports a batch mode. This can be used to run several estimates sequentially as one job, or to import predefined settings from a file. This feature is invoked using the **Run Script** button on the **Control Panel**, which processes script lines from the selected script file. The lines in a script file should have the format:

control_label [Parameter List]

where **control_label** defines the control to be set, and **[Parameter List] is a list of required parameters.** Entirely blank lines are ignored, and commenting is allowed (see below). The **control_label** matches the control labels used in the PRT interface, with underscore symbols replacing spaces, and all letters in lower case. The number of parameters used depends on the **control_label**, and are separated by blanks. No quotes are used, and no parameter can contain an embedded blank (so file names should be selected with this in mind).

A list of all available **control_labels** follows, with the required parameters indicated. The following conventions are used.

Any script line starting with ***** is treated as a comment. A space should separate '*' from any subsequent text.

The **control_label go** has the effect of pressing the **Go** button, using whatever control settings are current. It appears by itself in a single script line. The **Output Memo** can be cleared or saved with the **control_labels clear memo** or **save_memo**. The **Output Grid** is saved by the control label **save_output_grid**.

Commands which read or write files require a **Filename** parameter. For example, the script line

open_genotype_chart c:\ProjectDirectory\example.dat

reads data from the indicated file into the Genotype Chart. If a directory path is not specified, the current directory is used.

A parameter **N** refers to an integer. This value will be flagged for error if it does not conform to the restrictions of the particular control. For example, the script line

maximum_partition_list_size 10000

copies the value 10000 into the **Maximum Partition List Size** control in the **Partitions Options** panel, and this becomes the current control setting.

Script lines of the form

[#] 2.4_Run_Script_button

control_label N(1) to N(2)

are generally used to specify row and column ranges on the **Genotype Chart** and **Allele Chart**.

Script lines of the form

control_label T/F

are used to specify checkbox settings. For example

output_genotype_error_screen T

checks the **Output Genotype Error Screen** checkbox in the **Output Options** panel. The value of the parameter may be any string starting with T (true) or F (false), and is not case sensitive.

Script lines of the form

control_label Index

specify options from itemized lists. The value **Index** starts at 1. For example

select_enumeration_algorithm 2

makes the **MSG-T** algorithm the currently selected item in the **Selection Enumeration Algorithm** panel.

In addition the control label **clear_sibling_group_sizes** clears the grid in the **Sibling Group Sizes** panel, while the script line

add_sibling_group_sizes N(1) … N(k)

adds the line **N(1) … N(k)** to the grid, following any nonempty lines already present.

Application Section	Command	Required Parameters
General	*	
	save_memo	Filename
	clear_memo	
	save_output_grid	Filename
	go	
Genotype Chart	clear_genotype_chart	
	rotate_genotype_chart	
	open_genotype_chart	Filename
	save_genotype_chart	Filename
	relabel_genotype_chart	

	genotypes_in_rows	N(1) to N(1)
	genotypes_in_columns	N(1) to N(1)
	indiv_labels_in_columns	N(1) to N(1)
	group_labels_in_columns	N(1) to N(1)
	estimated_sg_labels_in_column	N
	numerical_validation_genotype_chart	T/F
	use_missing_value	T/F
	missing_value_code	N
Allele Chart	clear_allele_chart	
	rotate_allele_chart	
	open_allele_chart	Filename
	save_allele_chart	Filename
	relabel_allele_chart	
	loci_in_rows	N(1) to N(1)
	allele_freq_in_columns	N(2) to N(2)
	numerical_validation_allele_chart	T/F
Task	task	Index
Bootstrap Options	number_of_bootstrap_replications	N
	condition_on_alleles	T/F
	condition_on_genotypes	T/F
	unconditional_bootstrap	T/F
Partition Options	maximum_partition_list_size	N
	max_sib_size_for_full_search	N
	reduced_search_deletion	N
Randomization	specify_random_seed	T/F
	random_seed_text	N
Genotype Error Iteration	number_of_error_iterations	N
	maximum_errors_per_locus	N
Select Enumeration Algorithm	select_enumeration_algorithm	Index
MSG Algorithm Options	use_suggested_settings	T/F
	minimum_msg_size	N
	number_of_msg_iterations	N
	min_msg_size_increment	N
Triplet Enumeration Options	unlimited_triplet_iterations	T/F
	maximum_triplet_iterations	N
Output Options	output_genotype_error_screen	T/F
	output_alternative_sg_assignments	T/F
	output_nonexcluded_hsgs	T/F
	output_nonexcluded_parents	T/F
	output_bootstrap_plot_xy_data	T/F
	output_msg_list	T/F

	use_labels_for_partition_summary	T/F
Simulation Type	single_trial n_repeated_trials simulation_type	T/F N Index
Sample Size	n_indiv_in_sample	N
Frequency Properties	number_of_loci alleles_per_locus	N N
Sibling Group Sizes	clear_sibling_group_sizes add_sibling_group_sizes	 N(1) … N(k)

[#]Section 3. Defining Application Tasks

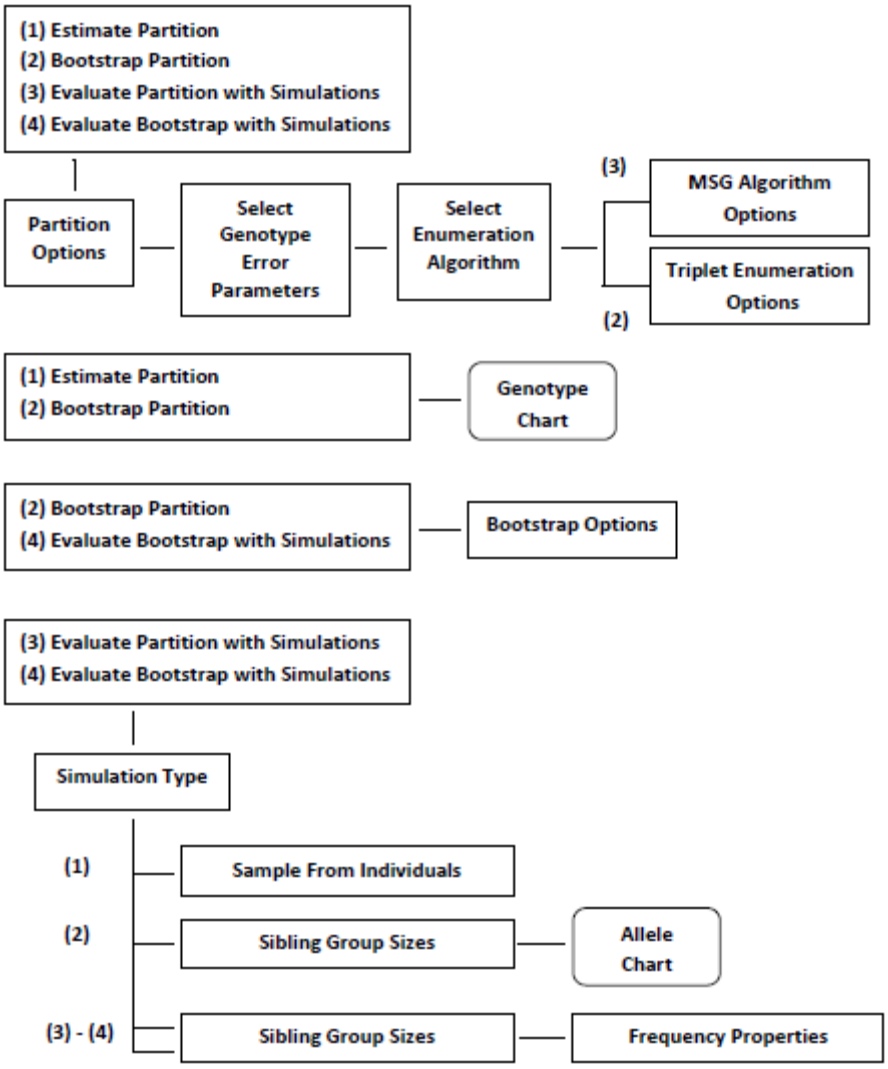
A job type is selected from the **Task** panel of the **Single Generation Tasks** tab sheet. Required parameters are specified elsewhere on the tab sheet, depending on the job. The job is activated by pressing the **Go** button. If genotype data or allele frequencies are needed, they are assumed to be present in the **Genotype Chart** or **Allele Chart**. The progress bars at the bottom of the **Control Panel** indicate the proportion of job completion. They may be turned off using the adjacent check boxes, which may reduce computation time (completion of a task will still be indicated in the **Output Memo**). The **Cancel** button will interrupt the job, in which case partial results are output when possible. A random number seed may be specified, which will be enforced when **Go** is pressed. Depending on the job, output can be found in the **Histogram Display**, **Bootstrap Coverage Display**, **Genotype Chart**, **Output Grid** or **Output Memo**.

Return to Table of Contents

[#] **Section_3._Defining_Application_Tasks**

[#]**3.1 Single Generation Tasks** tab sheet

The required panels depend on the task selected. The following four flowcharts indicate which panels are required based on the selected tasks and choices made on subsequent panels (tasks will appear in more than one flowchart).



Return to Table of Contents

[#] **3.1_Single_Generation_Tasks_**tab_sheet

[#]**3.1.1 Single Generation Tasks** tab sheet (Algorithm Details)

Task panel

One of four job types must be selected.
(1) Estimate Partition. This is a basic function among *single generation tasks*. It estimates a partition based on data in the **Genotype Chart**.
(2) Bootstrap Partition. This task estimates a confidence set of partitions using a bootstrap procedure based on the partition distance.
(3) Evaluate Partition with Simulations. With this task the user may estimate the accuracy of a partition algorithm by simulating genetic data from known partitions.
(4) Evaluate Bootstrap with Simulations. This is similar to task (3), except that the accuracy of the bootstrap procedure is evaluated.

Select Enumeration Algorithm panel

(1) MSG-G (graph based)
(2) MSG-T (triplet enumeration)
(3) MSG-MG (large SG modification)

There are two available methods of producing MSG lists: the MSG algorithm (MSG-G) and triplet enumeration (MSG-T). The MSG-MG Algorithm is a modification of the MSG algorithm designed for larger sibling groups. The difference is in the computation advantage. The MSG-G or MSG-MG algorithm is preferable with large sibling groups, while MSG-T is preferable with smaller groups. See section **1. Overview** for as discussion of the properties of these algorithms. Apart from the relative computational advantages, they function in the same way, and MSG-G and MSG-T will give identical output when used without modification.

MSG Algorithm Options panel

Both the MSG and MSG-MG algorithms may be thought of as belonging to a single class of algorithms. First note that the MSG algorithm may accept a threshold such that no MSG of smaller size will be output, and this may usually be done with significantly reduce computation time. The threshold may be any number larger than 2. This forms the basis of an iterative algorithm. First apply the algorithm using a large threshold T_1. Use the output to form a partial FSG partition. Apply the algorithm again using a smaller threshold T_2 after remove the previous constructed FSGs from the data set. This may be repeated until a full partition is formed. The following parameters must then be specified.

Minimum MSG Size. The minimum size of MSG to include in the list. This value need not be smaller than 3 (smaller values will be internally reset to 3). Otherwise, the algorithm will not include any

[#] **3.1.1_Single_Generation_Tasks_**tab_sheet_(Algorithm_Details)

MSG smaller than this value. For an iterative procedure this value represents the initial lower bound.

Number of MSG Iterations. The number of iterations to be used in an iterative implementation.

Min MSG Size Increment. When **Number of MSG Iterations** is greater than one this value specifies by what amount the lower bound is decreased after each iteration (starting from **Minimum MSG size**). Ideally, the three values in this panel are chosen so that the final lower bound is 3 or less (additional iterations after this point are ignored).

When using option **(1) MSG-G (graph based)** the **Minimum MSG Size** defaults to 3 and **Number of MSG Iterations** to 1 irrespective of the values currently entered into these fields (the **Min MSG Size Increment** is not used). When using option **(3) MSG-MG (large SG modification)** these parameters must be specified. There are two ways to do this. If the **Use Suggested Settings** check box is checked the program will set:

Minimum MSG Size	**= number of individuals in sample**
Min MSG Size Increment	**= 10**
Number of MSG Iterations	**= 2 + (Minimum MSG Size – 4) div 10**

These values will lower the threshold from the largest possible value to 3.

If the **Use Suggested Settings** check box is not checked, the program will use values entered into the **MSG Algorithm Options** panel.

Triplet Enumeration Options panel

This algorithm enumerates all MSGs by first enumerating all FSG triples, then iteratively merging the current FSG list until no further merges are possible. This method is computationally expensive when large sibling groups exist. Normally, this algorithm is used with the **Unlimited Triplet Iterations** option checked. To test the feasibility of the procedure, uncheck this option, then specify **Maximum Triplet Iterations**. This specifies the maximum number of iterations, and the size of the computation can be estimated from the output, which includes the size of the largest FSG found so far.

Partition Options panel

These options control the conversion of the MSG list to an actual partition, using the algorithm proposed in Almudevar and Field (1999). The *partition list* represents a type of workspace for the process, and is bounded by default to a length **Maximum Partition List Size** of 3000. Greater computational efficiency can be attained using a smaller value (say 500), but too small a value will compromise a satisfactory completion of the process.

Part of the process involves considering subsets of MSGs. A bound of **Max Sib Size for Full Search** is specified. For MSGs of size at or below this bound all subsets are considered. Above this bound all subsets obtainable by removing at most the number of individuals specified in the **Reduced Search Deletion** field (maximum 2). Therefore, computation time is reduced by lowering either of these values.

Genotype Error Iteration panel

PRT will attempt to flag genotypes as possibly in error. These are replaced with the current missing value, and the partition re-estimated (so please ensure that the **Missing Value Code** in the **Genotype Chart** doesn't conflict with actual allele values). To use this feature, enter two values in the **Genotype Error Iteration** panel (leaving both values equal to 0 turns off this feature). The **Maximum Errors Per Locus** value represents the maximum number of genotype errors permitted per individual. Values of 1 or 2 worked well in initial tests. In general, this value should reflect the anticipated error rate and the number of loci used. This feature can be much more effective when several iterations are permitted. A value of 5-10 for **Number of Error Iterations** worked well in initial tests. The iterations will stop if no genotype errors are flagged, so it is better to overestimate than underestimate this parameter.

Bootstrap Options panel

Three options are available for generating the sampling distribution for the bootstrap (conditional on alleles or genotypes, or unconditional). The algorithm returns a sampling distribution of the partition distance between the estimated partition and the true one. This is given numerically in the **Output Memo**, and graphically in the **Histogram Display**. More than one option may be specified, in which case only the last checked form of the bootstrap is displayed in the **Histogram Display**. Any attempt to uncheck all boxes will force **Condition on Alleles** to be checked as the default. See Almudevar (2001) for more detail.

Randomization panel

Simulations are used for tasks (2), (3) and (4) (the partition algorithm itself is deterministic). By default, the application uses the internal date and time to obtain a random generator seed. Optionally (by checking the **Specify Random Seed** box) the user may specify the **Random Seed**, so that jobs are reproducible. The random number is a signed 32 bit integer in the range -2147483648 ... 2147483647. The random seed is implemented whenever **Go** is pressed.

Output Options panel

Optionally, an auxiliary sibling structure information may be output with a partition:

Output Genotype Error Screen. If this option is checked then, following the partition estimate, each individual is checked to see if it can form a FSG with each other SG at all loci except exactly one. Each instance is reported (individual, SGs involved, locus and genotype which prohibit FSG status).

Output Alternative SG Assignments. If this option is checked then, following the partition estimate, each individual is checked to see if it can form a FSG with each other SG. Each instance is reported (individual, SGs involved).

Output nonexcluded HSGs. If this option is checked, then each pair of FSGs in the estimated partition are checked for compatibility as a HSG pair.

Output nonexcluded parents. If this option is checked, then a list of all nonexcluded parent pairs are output for each FSG in the estimated partition.

Output bootstrap plot XY data. If this option is checked, and the **(4) Evaluate bootstrap with simulations** option of the **Task** panel is selected, then the X-Y points needed to draw the bootstrap evaluation plot shown in the **Bootstrap Coverage Display** tab sheet are output to the **Output Memo** window.

Output MSG list. If this option is checked the full list of MSGs is output to the **Output Grid** window of the **Pedigree Reconstruction Tools** window. When the modified MSG algorithm is used, successive thresholded MSGs are appended into a single list.

Use labels for partition summary. If this option is checked, partitions which are output in the **Output Memo** window are labeled using the IDs provided in the **Genotype Chart** (if available).

Return to Table of Contents

[#]**3.1.2 Single Generation Tasks** tab sheet (Simulation Details)

Simulation Type panel

When task (3) or (4) is selected genotype data is random generated for the purpose of evaluating the partition algorithm or the bootstrap procedure. If a single simulation is selected (check **Single Trial**), then a single genotype sample is generated and written to the genotype chart (so that it may be saved for further analysis). This feature in not available for simulations generated by option **(1) Sample from Individuals** (see below). In addition, the output is of the same form as for tasks (1) or (2) respectively. The motivation is to provide greater scrutiny of a single replication. Alternatively, multiple replications may be selected, resulting in an aggregate estimate of the procedures accuracy.

There are four methods of generating simulated genotype data.

(1) Sample from Individuals from Genotype Chart. A data set is read directly into the Genotype Chart, which is assumed to contain Group ID fields. A genotype data sample is simulated by a random selection of individuals from this data. The sample size is given in the **Sample from Individuals** panel (see below).

(2) Use Allele Frequencies from Allele Chart. A partition is defined in the **Sibling Group Sizes** panel below. Genotypes for parents implied by the partition are generated using allele distributions which have been read into the **Allele Chart**, from which sibling genotypes are then simulated.

(3) Uniform Allele Frequencies. Similar to option (2), except that the allele distributions are assumed to be uniform based on parameters in the **Frequency Properties** panel (see below).

(4) Zipf Allele Frequencies. Similar to option (2), except that the allele distributions are assumed to be Zipf distributed based on parameters in the **Frequency Properties** panel (see below).

Sample from Individuals panel

N Indiv in Sample specifies the number of individuals in a simulated sample when option **(1)** is selected from the **Simulation Type** panel.

Frequency Properties panel

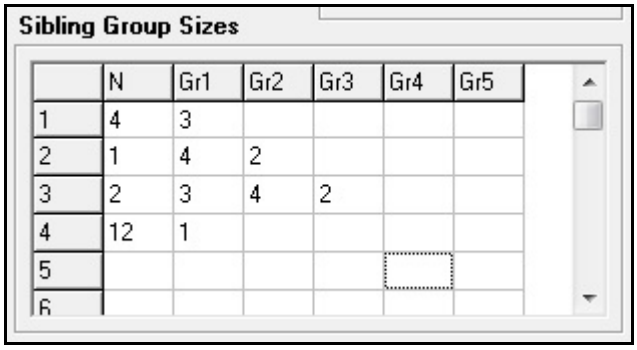
When option **(3) or (4)** is selected from the **Simulation Type** panel the number of loci, and the number of alleles per loci are specified by **Number of Loci** and **Alleles per Locus**.

[#] **3.1.2_Single_Generation_Tasks_**tab_sheet_(Simulation_Details)

Sibling Group Sizes panel

When option **(2)**, **(3) or (4)** is selected from the **Simulation Type** panel the partition structure is specified in this panel. Half sibling structure is supported. Each row may represent, for example, a specific mother, then each number in the row represents the number of offspring from a distinct father. The first column N specifies how many times the structure is to be repeated (repetitions are assumed to be unrelated).

For example, consider the following panel entry:



Row 1 represents 4 unrelated sibling groups of size 3. Row 2 represents a half sibling structure (one common parent) consisting of 2 sibling groups of sizes 4 and 2. Similarly, row 3 represents a half sibling structure consisting of sibling groups of sizes 3, 4 and 2. This structure is repeated twice, so that row 3 introduces a total of 8 unrelated parents. Row 4 represents 12 unrelated individuals. Empty rows (here, 5-100) are ignored.

Return to Table of Contents

[#]**3.2 Histogram Display** tab sheet

Several job types give as output a distribution of a partition distance error. In such cases this distribution is given numerically as a cumulative frequency distribution in the **Output Memo**, and as a histogram in the **Histogram Display** tab sheet. This applies to the bootstrap distribution produced by job type **(2) Bootstrap Partition** and job type **(4) Evaluate Bootstrap with Simulations** (when the **Single Trial** option is selected in the **Simulation Type** panel), and to the estimated partition accuracy distribution produced by job type **(3) Evaluate Partition with Simulations** (when the **Single Trial** option is not selected in the **Simulation Type** panel).

Return to Table of Contents

[#] **3.2_Histogram_Display_**tab_sheet

[#]**3.3 Bootstrap Coverage Display** tab sheet

For job type **(4) Evaluate Bootstrap with Simulations** (when the **Single Trial** option is not selected in the **Simulation Type** panel) the objective is to compare the actual confidence set coverage to the nominal coverage. Ideally, a coverage plot of the two quantities will lie along the identity. The procedure is sometimes conservative, and so the plot will often lie above the identity over some region. The procedure can be judged to be accurate when the plot lies close to the identity in the 0.9 to 1.0 range.

The X-Y points needed to draw the plot can be optionally written to the **Output Memo** window (see **Output Options** panel).

Return to Table of Contents

[#] **3.3_Bootstrap_Coverage_Display_**tab_sheet

[#]**3.4 Genotype Chart** tab sheet

The **Genotype Chart** stores genetic data, providing basic spreadsheet functionality. The cell entries are either alleles or labels (at the individual or the group level). Alleles are assumed to be nonnegative integers, while labels may be any text. It is assumed that genotypes are in consecutive columns, with two columns (of single alleles) representing a single locus.

Data input. Data may be read into the grid from a file (via the **Open** button). In this case, commas or spaces are accepted as a field delimiter (labels should therefore not contain spaces). The data may also be input manually, or edited at any time.

Numerical validation. When this option is checked only nonnegative integers may be entered.

Missing value code. A missing allele may be represented by a missing value code. If this is required then the **Use Missing Value** box should be checked. In this case, a missing value must be represented by an integer, which should be entered in the **Missing Value Code** text box.

Column definitions. The type of column entries must be correctly specified in the text boxes labeled **Genotypes in columns:**, **Indiv labels in columns:**, and **Group labels in columns:**. When only a single column is needed both text boxes should be filled (that is, columns "2 to 2" indicate entries in column 2). Individual labels are for reference only, and are not required. Group labels are needed when simulation type **(1) Sample from Individuals from Genotype Chart** is selected in the **Simulation Type** panel. If the text box labeled **Estimated SG labels in column:** has a nonzero entry then any partition estimate will be output to this column in the form of a sibling group index. This column is added by default when data is imported, but can be removed.

Row definitions. If only a subset of the rows are to be used, these are indicated by the text boxes labeled **Genotypes in rows:.** This will be useful if the file contains column headers. If all rows are to be used, these textboxes must indicate this. These values are set when a file is read into the grid, and may then be changed if needed.

Miscellaneous functions. The **Clear** button erases all entries. The **Rotate** button transposes the grid entries. This is useful if a file stores genotype entries by row instead of column. The **Open** and **Save** buttons are used to load and save grid entries using standard dialog boxes. The grid sheet displays column labels indicating the type and index of column entries. This is refreshed using the **Relabel** button. The number of rows and columns, and the column width, can be adjusted using the **Rows**, **Columns** and **Col Width** spin edit boxes.

Return to Table of Contents

[#] **3.4_Genotype_Chart_**tab_sheet

[#]**3.5 Allele Chart** tab sheet

The **Allele Chart** stores allele frequencies, providing basic spreadsheet functionality. The cell entries are real numbers, and each row represents one locus, in consecutive order. The rows will, in general, not contain the same number of entries. Allele frequencies are only used to simulate data, so need not be labeled. The frequencies will be normalized to sum to one, so ratio style entries such as "1, 1, 5, 20" may be input.

Data input. Data may be read into the grid from a file (via the **Open** button). In this case, commas or spaces are accepted as a field delimiter (labels should therefore not contain spaces). The data may also be input manually, or edited at any time. It is assumed that one line in a file represents one locus.

Numerical validation. When this option is checked only positive numbers may be entered.

Column definitions. The range of columns containing the frequencies are entered in the text boxes labeled **Allele freq in columns:**. This will be set when a file is read, and should be changed if, for example, a column contains labels.

Row definitions. If only a subset of the rows are to be used, these are indicated by the text boxes labeled **Loci in rows:.** This will be useful if the file contains column headers. If all rows are to be used, these textboxes must indicate this. These values are set when a file is read into the grid, and may then be changed if needed.

Miscellaneous functions. The **Clear** button erases all entries. The **Rotate** button transposes the grid entries. This is useful if a file stores genotype entries by row instead of column. The **Open** and **Save** buttons are used to load and save grid entries using standard dialog boxes. The grid sheet displays column labels indicating the type and index of column entries. This is refreshed using the **Relabel** button. The number of rows and columns, and the column width, can be adjusted using the **Rows**, **Columns** and **Col Width** spin edit boxes.

Return to Table of Contents

[#] **3.5_Allele_Chart_tab_sheet**