User Manual for Mendel's Accountant

Last updated on October 22, 2008

Contents:

Welcome to *Mendel's Accountant*.

This website provides an introduction to the program *Mendel's Accountant*. You may download the program at http://sourceforge.net/projects/mendelsaccount.

Mendel's Accountant (MENDEL) is an advanced numerical simulation program for modeling genetic change over time and was developed collaboratively by Sanford, Baumgardner, Brewer, Gibson and ReMine.

MENDEL is a genetic accounting program that allows realistic numerical simulation of the mutation/selection process over time. MENDEL is applicable to either haploid or diploid organisms, having either sexual or clonal reproduction. Each mutation that enters the simulated population is tracked from generation to generation to the end of the experiment - or until that mutation is lost either as a result of selection or random drift. Using a standard personal computer, the MENDEL program can be used to generate and track millions of mutations within a single population.

MENDEL's input variables include such things as mutation rate, distribution specifications for mutation effects, extent of dominance, mating characteristics, selection method, average fertility, heritability, non-scaling noise, linkage block properties,

chromosome number, genome size, population size, population sub-structure, and number of generations.

The MENDEL program outputs, both in tabular and graphic form, provide several types of data including: deleterious and beneficial mutation counts per individual, mean individual fitness as a function of generation count, distribution of accumulating mutation effects, selection threshold over time, distribution of linked mutation effects, fitness distributions before and after selection, and allele frequencies.

MENDEL provides biologists with a new tool for research and teaching, and allows for the modeling of complex biological scenarios that would have previously been impossible.

MENDEL's Basic Principles of Operation.

The actual design and code features of MENDEL are described in detail elsewhere *[Sanford et al., 2007. SCPE 8(2): 147-165 - also available on this web site]*. Following is a simple outline of how the program operates.

1. Based on user input, Mendel creates a virtual population with the specified number of individuals, mean reproduction rate, mating characteristics, and possible population sub-structure. Also specified are the genome size, number of chromosomes, and linkage dynamics.
2. Based on user input, Mendel creates a pool of potential mutations with precisely specified characteristics including: range and frequency distribution of mutation effects, ratio of recessives to dominants, and fraction of beneficial mutations.
3. Mutations are selected randomly from the pool of potential mutations and are assigned randomly to the new offspring in each generation, based upon the specified average mutation rate (Poisson distribution).
4. Individual genetic fitness values are calculated based upon each individual's total mutation inventory. Individual genetic fitness is defined as 1.0, adjusted by the positive and negative effects of all its mutations. To obtain phenotypic fitness the genetic fitness is modified using the specified heritability to account for non-heritable factors such as variations in the environment.
5. Based on phenotypic fitness values, Mendel applies selection of a specified type (probability, truncation, etc.) to eliminate a specified fraction of the individuals from the mating pool. The fraction removed is determined from the fertility of the population (average offspring per female). Offspring in excess of two per female represent the surplus that is eliminated via selection based on individual phenotypic fitness scores. Unless otherwise specified, population size is kept constant by selecting away the entire surplus population each generation.
6. Gametes are extracted from those individuals that survive the selection process and reproduce. These gametes are generated based upon a sampling of the linkage

blocks within each reproducing individual. Reproducing individuals are paired off randomly, and their gametes are fused to create the next generation of individuals.

Steps 1-6 are repeated, for the specified number of generations. Output reports and plots are updated at regular intervals.

Computer Requirements.

MENDEL was designed to run on a wide variety of computer platforms and is currently supported on Microsoft Windows and Fedora Linux systems. MENDEL requires the following support software: Apache Web Server, Perl, and Gnuplot. For the Linux version two additional programs are required: Torque Resource Manager (or open PBS) and MPICH2 (or an MPI variant). A recommended hardware requirement is a minimum of 512MB of RAM. Although MENDEL can run with less RAM, the size of a run (largely determined by population size and mutation rate), is limited by the available memory. The Linux version is currently the superior version because the Windows version has more limited capabilities. However, the Windows version should be very useful in teaching environments. For both versions, performance is significantly enhanced when a dual processor (or dual core) machine is used, because the graphical user interface does not need to compete for resources with the Mendel program itself. A stand-alone dual core processor machine or a Linux cluster is therefore strongly recommended. For additional documentation on setting up MENDEL on Linux, please download the "MENDEL Linux How-to" from http://mendelsaccountant.info.

Downloading Instructions.

MENDEL can be downloaded by clicking through a series of green download buttons at http://sourceforge.net/projects/mendelsaccount. You must specify whether you want the Linux or Windows version. If "Downloading" appears on the screen but the download does not appear to be proceeding – click "this direct link" which is shown in blue. Once downloaded, carefully follow the instructions for installation. The Windows version is ready to use once it is installed. However, the Linux version cannot be used until five supporting programs have been downloaded (Apache Web Server, Gnuplot, Perl, Torque Resource Manager, and MPICH2). Because it is more difficult to make the Linux version operational, we will be happy to assist potential users in this process. To make it easier for potential users to evaluate Mendel, we are setting up several locations where Mendel can be operated by approved users on one of our own computer clusters via remote internet access. In this case no downloading or installation is required – all that is needed is a password (visit http://hs84.nysaes.cornell.edu and click "Sign up" – it may take between 24 to 72 hours for your account to be approved.

Opening and operating MENDEL.

MENDEL is controlled through its own web-based graphical user interface. Therefore, to access MENDEL, one must open a web browser. (*Google Chrome*, *Apple Safari* or *Internet Explorer* are recommended for the Windows version, and *Konqueror* for the Linux version. However, *Firefox* works in both Windows and Linux but is not recommended as it does not support all of the features in the parameters interface.).

1. To open the program one must click on the icon (Windows), or one can enter the following address in the browser -- http://127.0.0.1:8080/mendel/v1.2.1/index.html, or select the website for Linux systems (e.g. http://myservername.com/mendel/v1.2.1/index.html). or select the local website for Linux systems (e.g. http://myservername.com/mendel).
2. To start a fresh run, simply press "Start", and begin to fill in the parameter options as desired (initially there are default input parameters in all boxes).
3. To start a run using most or all of the parameters from a previous run, enter the previous run's case ID in the box (panel of options on left of screen), and then press "Start". The input options will now have the parameters from that earlier run, which can then be modified as desired prior to execution.
4. Once the parameters have all been entered, press "submit". Review the selected options, and if necessary use the browser "back" button to go back and modify the parameters. Check to see if the required memory is greater than the available memory before pressing the "execute" button. If the "required" memory is greater than the available memory, in many cases the run will still be completed successfully, but in some cases it will fail to execute or will "crash" before the run is complete.
5. If the run has been successfully activated, a "Run Status" report will come up, specifying the run's number and it time of initiation, and its progress.
6. During the run, one can periodically examine the output file.
7. During the run, one can periodically refresh and examine the output figures.
8. During the run one can review the "Input Parameters", or check the "Run Status".
9. If desired, one can open "Run Status", and terminate a run by selecting the current run and then selecting "Stop".
10. During or after a run, one can click "Label" and add comments to a run, for future reference.

MENDEL Input Parameters.

MENDEL's default input values are parameters that might apply to a small human population. Alternatively, template parameter settings for a small yeast population are also provided (these yeast parameters are only educated guesses - we have not researched what parameters would be most reasonable for yeast). These default parameters are readily altered to be applicable to any haploid or diploid population having either sexual or clonal reproduction. This includes self-fertilizing plant species and species with very

small genomes. MENDEL's input parameters have been separated into two categories, 'basic' and 'advanced'.

Basic Parameters:

The basic input parameters include the variables that most obviously affect mutation accumulation, and are most frequently changed. This parameter set is useful for initial familiarization with Mendel, and for training beginning students.

1. The first basic parameter is case ID. Since MENDEL allows each run to be saved, each experiment needs to be identified for the purpose of data retrieval. The ID must be alphanumeric and must have exactly 6 characters. We recommend users begin their case ID with their own initials.
2. The second basic parameter is the average number of new mutations per individual. In humans, this number is believed to be approximately 100. The mutation rate can be adjusted to be proportional to the size of the *functional* genome. Thus if only 10% of the human genome actually functions (assuming the rest to be biologically inert), then the mutation rate would be reduced from 100 to just 10. Rates of less than 1 new mutation per individual are allowed - including zero. The human default value is 10 new mutations per individual per generation.
3. The third basic parameter is the ratio of beneficial versus total mutations. While some sources suggest this number might be as high as 1:1000, most sources suggest it is more realistically about 1:1,000,000. The default setting is 1:100,000. For studying the accumulation of only deleterious mutations, the number of beneficials can be set to zero.
4. The fourth basic parameter is the number of offspring per female. Since population size in Mendel is usually constant, this variable defines the maximum amount of selection. There must be an average of at least 2 offspring per female for the population to maintain its size and avoid rapid extinction. Except where random death is considered (see advanced parameters), all offspring in excess of 2 are removed based upon phenotypic selection. The default value is 6 offspring per female.
5. The fifth basic parameter is the desired population size. This is the number of reproducing adults, after selection. This number is normally kept constant, except where fertility is insufficient to allow replacement, or where certain advanced parameters are used. For smaller computer systems such as PCs, population size must remain small (100-1000) or the program will quickly run out of memory. The default value is 1,000, since population sizes smaller than this can be strongly affected by inbreeding and drift. We find increasing population size beyond 1000 results in rapidly diminishing selective benefit.
6. The sixth basic parameter is the number of generations the program should run. The default is 500 generations. If there are too many generations specified, smaller computers will run out of memory because of the accumulation of large numbers of mutations, and the experiment will terminate prematurely. This problem can be mitigated by tracking only the larger-effect mutations (see

advanced computation parameters). The program also terminates prematurely if fitness reaches zero or if the population size shrinks to just one individual.

Advanced Parameters:

Advanced Parameters: The advanced parameter settings are for knowledgeable researchers and more advanced students, and should be set to apply to a specific species or circumstance. The advanced parameters are distributed among the general categories of: Mutation, Selection, Population Biology, and Computation.

Advanced Mutation Parameters

1. Parameters shaping the distribution of deleterious mutations. Deleterious mutations in the natural world typically range from a few rare lethals to a large number of nearly-neutral mutations (entirely neutral mutations have already been discounted - see above section on mutation rate). It is widely agreed that the distribution of mutational effects is characterized by an exponential-like function. Mendel uses a generalized exponential function, called the Weibull function, to generate its distribution of mutation effects ranging from 1 (lethal) down to nearly 0 (neutral). See Sanford et al., SCPE 8(2) p.147-165 (available on this website), for the mathematical formula that describes Mendel's mutation effect distribution, and also for an explanation of the difference between "fitness effect" and "selection coefficient". The exact shape of the mutation effect distribution produced by MENDEL can be controlled precisely, using three variables:
    a. Functional haploid genome size - The relative abundance of near-neutrals depends, in part, upon genome size. A simple viroid with a genome size of just 100 nucleotides may have very few or even zero near-neutral nucleotide positions. In such a viroid, we might assume that a slightly deleterious mutation would generally reduce genomic information by (very roughly) about one part in a hundred. Alternatively, a human genome of 3 billion will have very many near-neutral positions, and a near neutral mutation might (very roughly) reduce genomic information by only one part in 3 billion. Therefore, it is obvious that the distribution of deleterious mutational effects must in some way be adjusted to account for genome size. An approximate yet reasonable means for doing this is to define the minimal mutational effect as being 1 divided by the functional haploid genome size. The result of this adjustment is that smaller genomes have "flatter" distributions of deleterious mutations, while larger genomes have "steeper" distribution curves. Because we are discounting all entirely neutral mutations, we must only consider the size of the functional genome, so we choose the default genome size to be 300 million (10% of the actual human genome size).

b. Fraction of mutations having "major effect" - Most mutations have an effect on fitness that is too small to measure directly. However, mutations do have measurable effects in the far "tail" of the mutation distribution curve. By utilizing the frequency and distribution of "measurable" mutation effects, one can constrain the most significant portion of the distribution curve as it relates to the selection process. For most species, there may not yet be enough data, even for the major mutations, to accurately model the exact distribution of mutations. When such data is not yet available, we are forced to simply estimate, to the best of our ability based on data from other organisms, the fraction of "major mutations". The default is 0.001.

c. Cut-off point for defining "major effect" - A somewhat arbitrary level must be selected for defining what constitutes a "measurable", or "major", mutation effect. MENDEL uses a default value for this cut-off of 0.10. This is because under realistic clinical conditions, it is questionable that we can reliably measure a single mutation's fitness effect when it changes fitness by less than 10%.

2. Parameters shaping distribution of beneficial mutations. The distribution of beneficial mutations should generally be a mirror image of the distribution of the deleterious mutations, except that the area under the distribution curve should be adjusted to reflect the proportionately lower number of beneficial mutations compared to deleterious mutations. Since the distribution of beneficials should be affected by genome size (as with deleterious mutations), it is useful to likewise define the minimal beneficial mutation effect as 1 divided by the functional haploid genome size. In addition, beneficials should have a reduced upper range, as described below.

   a. Maximal beneficial mutation effects – A realistic upper limit must be placed upon beneficial mutations. This is because a single nucleotide change can expand total biological functionality of an organism only to a limited degree. The larger the genome and the greater the total genomic information, the less a single nucleotide is likely to increase the total. Researchers must make a judgment for themselves of what is a reasonable maximal value for a single base change. The MENDEL default value for this limit is 0.001. This limit implies that a single point mutation can increase total biological functionality by as much as 0.1%. In a genome such as man's, assuming only 10% of the genome is functional, such a maximal impact point mutation might be viewed as equivalent to adding 300,000 new information-bearing base pairs each of which had the genome-wide average fitness contribution. Researchers need to honestly define the upper limit they feel is realistic for their species. However it should be obvious that, in all cases, the upper limit for beneficial mutation effects ought to correspond to a very small fraction of the total genomic information (i.e. a small number relative to one).

   b. Number of initial beneficial mutations – This parameter lets the researcher begin a run with a specified number of pre-existing beneficial mutations already in the population. These pre-existing mutations will follow the

specified beneficial mutation distribution. These alleles will all begin as single-copy mutations within the population. *(Caution - at present, using pre-specified mutations can cause errors in some of the output plots).*

3. Parameters involving recessive and dominant mutations. It is widely agreed that most mutations are recessive, while a small fraction are dominant. However, because modeling recessives and dominants can become computationally more intense, and because the output figures can become hard to read when plotting both dominant and recessive mutations, the default setting is co-dominance. This means that all mutations behave additively (a heterozygote will always have half the effect of a homozygote). However, for greatest realism, the majority of mutations should be made recessive, with a minority being dominant by default, as described below:
    a. Fraction of mutations recessive – This parameter simply specifies the percentage of mutations that are recessive. The default is 0.0%. If set to 80%, then 80% of mutations are recessive, so the remaining 20% will automatically be made dominant.
    b. Recessive expression in heterozygotes – It is widely believed that recessive mutations are not completely silent in the heterozygous condition, but are still expressed at some low level. Although the default is 0.0%, a reasonable setting would be 5%.
    c. Dominant expression in heterozygotes - It is widely believed that dominant mutations are not completely dominant in the heterozygous condition, but are only expressed only at some very high level. Although the default is 50%, a reasonable setting would be 95%.
4. Fraction multiplicative effect. When there are two or more mutations within an individual, the effects of these multiple mutations must be combined. The most straightforward way to do this is additively, by just adding up the effects of all the deleterious and beneficial mutations within an individual, and adjusting original fitness (initially 1.0) by that net amount. Alternatively, one can adjust fitness by multiplying the fitness (initially 1.0) by the net effect of each mutation (the net effect of a single mutation would be one minus the fitness effect of that mutation). This multiplicative method is quite commonly used in population genetics, although in our experience it seems inadequate when modeling reality - since no amount of deleterious mutation can drive fitness completely to zero. The reason for this is that in the multiplicative model, each additional deleterious mutation has less and less effect on absolute fitness. For this input parameter, the researcher can select an all additive model (0.0 multiplicative = default), or an all multiplicative model (1.0, no additive component), or a mixed model having any intermediate value between 0 and 1.0. MENDEL's default setting is the simple additive method. A third way to combine mutational effects is to use a synergistic epistasis model (see #6 below).
5. Special case: All mutations are given an identical fitness effect. It is sometimes of theoretical interest to model what would happen if all deleterious mutations had an equal effect and all beneficials had an equal effect. If this option is selected, one needs to specify those constant values for both deleterious and beneficial mutations. When this special case is chosen, it overrides the actual mutation

distribution parameters otherwise specified. Selection is consequently based essentially upon each individual's total mutation count.

6. Special case (Linux only): synergistic epistasis (SE). As described above (#4), it is our opinion that the most realistic way to combine mutations when calculating their net fitness effect within a given individual is simply to add up their effects (the additive model). An alternative method for combining mutation effects is to multiply their affects together such that each new deleterious mutation has less and less effect (the multiplicative model). In some very rare instances, researchers wish to invoke the exact opposite – wherein each additional mutation has a greater and greater effect. This treatment, called *synergistic epistasis,* was conceived as a theoretical mechanism that might in principle halt the accumulation of deleterious mutations. The basic idea is that deleterious mutations might act in many cases to compound the effects of one another. Apart from selection, this would, of course, produce accelerated degeneration. However, with selection, there might be more intense selection against individuals possessing more mutations than average. It was reasoned that this might conceivably halt mutation accumulation.

   Interactions between genetic units certainly happen (which is termed epistasis), and Mendel allows us to model a special generic case of negative epistasis called synergistic epistasis (SE). In our SE treatment, each mutation's deleterious fitness effect becomes amplified more and more (the SE penalty) as the individual's total mutation count increases. The user must specify exactly how much the SE penalty increases as mutation count increases. In addition, the user must specify what fraction of all SE interactions are within the same linkage block (the effects of such interactions will be fully heritable), versus the fraction of SE interactions which are between unlinked mutations (the effects of which will not be heritable).

   Specifying the fraction of mutation/mutation interactions which are linked

   In general, one expects synergistic effects to be larger and more frequent when the interacting mutations are in close proximity within the genome, for example, within the same gene. Any such SE effect from a physically linked interaction should be inherited in the normal Mendelian way, just as if it was an extra deleterious mutation within that linkage block. This type of linked SE interaction is modeled by simply adding a heritable fitness effect value (the SE penalty), to the relevant linkage block. This SE penalty will be transmitted 100% to the progeny.

   On the other hand, specific SE effects from non-linked interactions are created and destroyed every generation - as recombination occurs. These interactions have temporary fitness effects that impact only one generation at a time. This type of epistasis is normally considered a type of noise, as far as the selection process is concerned. However in the SE model, as mutation count per individual increases, there is a net increase in the number of negative mutation-to-mutation

interactions, creating an escalating deleterious effect for high mutation-count individuals - potentially enhancing selection against such high mutation-count individuals. Only the non-linked SE interactions have any potential for enhancing selection efficiency in this way.

Therefore, the first input parameter governing the SE treatment specifies the fraction of non-linked SE interactions relative to those from non-linked interactions, (linked_mutn_se_fraction). This parameter can be set anywhere from 0 (no interactions are linked) to 1.0 (all interactions are linked).

Specifying the maximal SE penalty

The second input parameter governing the SE treatment specifies the maximal magnitude of the SE penalty (se_scaling_factor). This parameter applies equally to both linked and unlinked interactions. It can vary from zero (no SE penalty), to an indefinitely large number. If set at 1 (the default setting) the maximal SE penalty will double the effect of each mutation when every site is mutant. If set at 1 million, the maximal SE penalty will amplify the effect of each mutation a million-fold under those same circumstances.

Advanced Selection Parameters

1. Rate of "Random Death". A certain fraction of any population fails to reproduce, independent of phenotype. This can be expressed as the percentage of the population subject to random death. This is a useful parameter conceptually, but the same effect can be obtained by proportionately decreasing the number of offspring/female.
2. Heritability. Because a large part of phenotypic performance is affected by an individual's personal circumstances (the "environment"), selection in nature is less effective than would be predicted simply from genotypic fitness values. Non-heritable environmental effects on phenotypic performance must be modeled realistically. MENDEL's default value for the heritability is 0.2. This implies that on average, only 20% of an individual's phenotypic performance is passed on to the next generation, with the rest being due to non-heritable factors. For a very general character such as reproductive fitness, 0.2 is an extremely generous heritability value. In most field contexts, it is in fact usually lower than this.
3. Non-scaling noise. If a population's fitness is increasing or declining, heritability (as calculated in the normal way), tends to scale with fitness, and so the implied "environmental noise" diminishes or increases as fitness diminishes or increases. This seems counter-intuitive. Some researchers may wish to model a component of environmental noise that does not scale with fitness variation. The units for this non-scaling noise parameter are based upon standard deviations from the initial fitness of 1.0. For simplicity, the default value is 0, but reasonable values probably exceed 0.01 and might exceed 0.1. The specified heritability and non-scaling noise both represent non-heritable variation, so non-scaling noise makes the effective heritability lower than the specified heritability. Mendel outputs the

standard deviation for genotypic and phenotypic fitness before and after selection, so effective ("realized") heritability can be calculated if the user so desires. The default value is "0" (no non-scaling noise).

4. Fertility declines as fitness declines. It is widely recognized that when fitness declines, fertility also declines. This in turn affects population surplus, which affects selection efficiency, and can eventually result in "mutational meltdown". To model this, we have included an option wherein fertility declines proportional to the square of the fitness decline. The resulting fertility decline is initially very subtle, but becomes increasingly severe as fitness approaches zero. The default value is "Yes", which means that fertility declines with fitness, especially as fitness approaches zero.
5. Type of selection. MENDEL's default mode for type of selection is probability selection, wherein the probability of reproduction is proportional to an individual's fitness ranking within the population. Two forms of probability selection are provided—classic and unrestricted. In classic (textbook) probability selection, rather counter-intuitively, strict proportionality (relative to the most-fit individual) can combine with high average fitness and mild selection (low reproductive rates) to cause reductions in fitness and relatively rapid extinction. In unrestricted probability selection, with certain combinations of average fitness and offspring/female, a range of the highest fitness values are "guaranteed" survival in order to maintain population size. To give the researcher maximal flexibility, we also provide an option where strict truncation selection (i.e. artificial selection) is employed. An intermediate option, involving a form of "broken-line" selection which we have designated "partial truncation" has also been added. With certain combinations of fitness distribution and offspring/female, this selection mode involves either truncation of the least fit individuals, guaranteed survival of the most fit individuals, or both - with probability selection acting on the balance of the population.

Advanced Population Parameters

1. Reproductive mode. Normal sexual reproduction is the default setting, but clonal reproduction can be specified. If clonal reproduction is selected, there is no recombination (overriding #4 below), and the genome is treated as one large non-recombining chromosome. There is no mating, and the same genome is transmitted from female to offspring, with each offspring then being assigned its own set of new mutations.
2. Ploidy Level. Diploidy is the default setting, but haploidy can be specified. In this special case, selection occurs during the haploid phase of the reproductive cycle, which makes all mutations "dominant" (always expressed 100%).
3. Fraction self-fertilization Certain plants and lower animals can self-fertilize. The percentage of self-fertilization (as opposed to out-crossing) can be set to range from the default value 0%) up to 100%. As this value increases, there is a strong increase in inbreeding and in the rate of mutation fixation. Consequently, recessive loci have a much stronger effect on overall fitness than normal.

4. Initial heterozygous alleles. Under certain conditions it is desirable to start a run with a set of initial heterozygous alleles (all individuals heterozygous at those specified loci). To do this one must specify the number of alleles, and their maximal combined impact.
    a. *Number of initial contrasting alleles:* This input lets the researcher begin a run with a specified number of initial contrasting alleles (heterozygous alleles), with a positive and negative allele at each contrasting locus in each individual. This gives an initial frequency of 50% for each allele, where each allele is co-dominant. This situation is analogous to an F_1 population derived from crossing two pure lines or two relatively uniform breeding lines of animals, and is very roughly analogous to natural crossing of two isolated populations in nature. This input allows investigation of the effect of factors such as environmental variability, type of selection, and percent selfing on the retention of beneficial alleles during segregation after a cross. Initial heterozygous alleles are not to be confused with initial beneficial mutations, which are represented as only a single copy in the initial population. Each contrasting locus is on a separate linkage block, such that the maximum number of initial contrasting loci is the number of linkage blocks. The contrasting alleles are assigned to linkage blocks equally spaced along the genome. With dynamic linkage (see input parameters), if the number of alleles specified is much less than the number of linkage blocks, the alleles will segregate completely independently of one another. If the number of loci specified is large relative to the number of linkage blocks (~ > 3 * number of chromosomes), segregation of these alleles will involve some degree of linkage. With static linkage, each allele will always separate independently.
    b. *Maximum total fitness increase*: The maximum total fitness increase is the amount the fitness which would be increased if all the positive alleles became fixed (homozygous in every individual). Realistically, this value would always be considerably less than 1. A value of 1 would potentially double the mean fitness ("yield" in plant breeding situations). Such a large potential increase would be larger than most situations encountered in nature or in plant or animal breeding. The actual fitness increase in the population will actually always be less than the maximum total fitness increase (unless selection moved all the positive alleles to fixation). If the number of initial contrasting alleles is no more than 10, each positive allele is assigned an equal fitness value representing the maximum total fitness increase divided by the number of initial contrasting loci. If the number of alleles is greater than 10, then each positive allele is assigned a random value from a distribution of values whose mean is the maximum total fitness increase divided by the number of initial contrasting loci.

       Outputs associated with initial contrasting alleles are the mean fitness contribution of positive contrasting alleles, mean frequency of positive contrasting alleles, and number of positive contrasting fixed or lost. The

effect and frequency of individual alleles is listed at the end of the output file.

A planned future improvement is to allow a user-specified degree of dominance in the initial heterozygous alleles. This would allow more direct investigation of the role of dominance in the expression of heterosis and inbreeding depression.

5. Dynamic linkage. Linkage has a major effect on selection, and must be modeled as accurately as possible. MENDEL's default mode involves dynamic linkage. This requires specification of the haploid chromosome number and assumes two random crossovers within each chromosome, at random locations between linkage blocks - every generation. Because tracking every linkage block can become computationally expensive, the number of linkage blocks must be limited (default = 1,000, min=1, max=100,000). The number of linkage blocks is evenly distributed over a user-specified haploid number of chromosomes (default=23). We also offer the researcher the option of a simpler model involving the specification of a fixed number of linkage blocks and fully randomized recombination between all linkage blocks each generation (no chromosome number is required).
6. Dynamic population growth (Linux only). By default Mendel uses a static population size. However, two options are provided to simulate dynamic population growth: (1) exponential growth model (e.g. Figure 1), and (2) carrying-capacity model. For the exponential growth model, two additional inputs need to be entered: *population growth rate* and *maximum population size*. The population growth rate parameter determines the percent growth rate per generation. A value of 1.0 represents static population size (no growth). To grow the population 2% per generation, enter the parameter 1.02 (note: one may need to manually convert published annual population growth rates to population growth per generation by using a formula such as 1.02^{20} = 1.48/generation – assuming a 20 year generation time). If the exponential growth model is selected, the "number of generations" input parameter (under *basic parameters*) is alternatively used to input the maximum population size (the number of required generations being unknown). Number of generations is no longer necessary, because the simulation will automatically shutdown when the computer runs out of memory or the maximum population size is reached. The population growth parameter may vary between 1.0 – 1.5. The user will notice for low population growth rates (e.g. 1.001) starting with small population sizes, the population will grow linearly until a sizeable enough population is formed, and then true exponential growth will occur. This initial linear phase is because population size in an integer number.

   Mendel's second population growth model is called "the carrying-capacity model". Wikipedia.org *(accessed October 7, 2008)* gives the following definition for carrying capacity: "The supportable population of an organism, given the food, habitat, water and other necessities available within an environment is known as

the environment's carrying capacity for that organism." The equation describing relating population growth to the environment's carrying capacity can be given as:

$$dN/dt = rN(K-N)/K$$

[Reference:Halliburton, Richard. Introduction to Population Genetics, Benjamin Cummings, 2003]. where *N* is the population size, *r* is the maximum reproductive rate of an individual, and *K* is the carrying capacity. This equation is solved numerically as:

$$N_{i+1} = N_i\,(1 + r\Delta t\,[1 - N_i/K])$$

Where *i* represents the generation number, $r\Delta t$ is a constant which represents the maximum reproductive rate of an individual times over a period of time (here about 20 years). Presently, this constant may only range between *0.0* to *1.0*.

7. Population sub-structure (Linux only). Perfectly random mating within a population probably never happens, especially in larger dispersed populations. MENDEL allows creation of multiple sub-populations, and allows the specification of the rate of migration between sub-populations, using three possible methods (one-way stepping stone, two-way stepping stone, and island).
8. Population bottlenecks. Population bottlenecks can dramatically affect mutation accumulation and mutation fixation. MENDEL allows the modeling of population bottlenecks. The researcher can cause a bottleneck to automatically begin after a specified number of generations, resulting in a specified reduction in population size, and ending after a specified number of bottleneck generations. The reduction of population size occurs immediately at the beginning of the bottleneck, by selecting a random sub-sample of the population. When the bottleneck ends, the original offspring number/female does not change, but half of the population excess (i.e. all offspring exceeding 2 per female) is used to increase population size, and half of the excess continues to be eliminated by selection. When the original population size is reached, normal selection is restored.

Advanced Computational Parameters

1. Tracking threshold. MENDEL can track every individual mutation. However, this may not be the best choice, especially with large populations and/or large numbers of generations. Hundreds of millions of mutations can accumulate within a virtual MENDEL population, causing computer operating speed to slow to a crawl, and eventually exceeding all available memory. In order to speed operation and allow larger experiments, MENDEL can track only those individual

mutations that are "potentially meaningful". Most mutational effects are so close to zero, that they can be classified as "extremely near-neutral". Such effects are so extremely small that they have no significant impact, even after accumulating to very high numbers for many, many generations. Such mutations may more practically be dealt with as follows: (1) they can be assumed to act in a co-dominant manner [JRB2] (such that their dominant/recessive status can be ignored); (2.) their effects can be pooled into their respective linkage block effects; and (3.) their number can be monitored using a "mutation counter". For the sake of computational efficiency, the researcher only needs to make a practical decision in terms of where the cut-off value should be for defining "extreme near-neutrals". Above this threshold, all mutations are still individually tracked in the usual way. MENDEL's default for this threshold is 0.00001. Where high speed and maximal use of memory is desired, the tracking threshold can even be set at 1.0. This results in zero tracking of individual mutations - all mutation effects are dumped into the appropriate haplotype, so tracking is only by haplotype. The advantage of this option is that much larger runs can be done much faster, given the same computing resource. The disadvantage is that the information normally found in figures 2, 3, and 5 will be lost and all mutations are made co-dominant by default.

2. Random number seed. At several stages within the MENDEL program, a random number generator is required. When an experiment needs to be independently replicated, the "random number seed" must be changed. If this is not done, the second experiment will be an exact duplicate of the earlier run.
3. Changing parameters over time. MENDEL allows a run to go for a specified number of generations, followed by data output, alteration of certain biological parameters, and resumption of the run. This can be done repeatedly, simply by choosing the commands "**allow this run to be re-started**" prior to a run, and then later "**restart new phase of run**…" prior to subsequent runs. Most parameters can be altered at restart, but population size and the number of linkage blocks must remain unchanged. (Caution: allowing restarts of large runs will save large amounts of data, which can rapidly fill available disk storage.).
4. Parallel processing (Linux only). To do larger experiments or to reduce the wall clock time for a job, or to simulate population sub-structure (tribes), MENDEL provides the option of running a job in parallel on multiple processors. In this case the number of processors must be specified. Plot option – when multiple processors are used to represent sub-populations (tribes), the user must specify if output plots should be show data the whole population or for each processor/tribe separately.
5. Batch or single high-memory processor (Linux only). In most cases, this should be set at "Batch". However, on one specific server, selecting "single high-memory processor" will cause the job to be run on a special compute node which has 16GB of memory.
6. Fortran verses C version (Linux only). Currently the Fortran version is used by default and is faster, and consumes less memory. However, due to the popularity of C-based languages, we have developed a C version which is under testing, and gives good agreement with the Fortran version. For doing parallel simulation (e.g. multiple tribes) the C version appears to work much better and is supported on

more servers than the Fortran version (the parallel version of Fortran is not supported on most servers).

MENDEL Output.

MENDEL routinely outputs a primary summary data file and a series of figures. Raw data output files are also saved.

Output summary data file: This primary summary data file involves an on-going real-time reporting of deleterious and beneficial mutation counts, mean fitness, fitness standard deviation, as well as other data. These summary statistics are reported for generations 1, 2, 3, 10, 20, and then every 20 generations.

Raw data files: These can be accessed separately through a marked box within each output figure.

Figure 1: This figure summarizes mutation counts per individual and average fitness - both plotted versus generation count. **Figure 1a** plots the number of both deleterious and beneficial mutations per individual. The scale for deleterious mutations is on the left and the scale for the beneficial mutations is on the right. **Figure 1b** shows average individual fitness (left scale) and the standard deviation for fitness (right scale), plotted versus generation count. The initial fitness is always assumed to be 1.0. For both 1a and 1b, the y-axis is self-scaling, since mutation counts and fitness can change dramatically over time.

Figure 2: This figure is designed to reveal how selection is altering the distribution of mutation effects in the population. Typically it shows that the frequencies in the population of mutations with larger effects are affected by the selection process much more strongly than the frequencies of mutations with small effects. Figure 2a is histogram plot that shows the mutation effect distribution for deleterious mutations. The x-axis uses a log scale to represent the magnitude of the mutation effects and ranges from lethal (-1.0) on the left to the minimal mutation effect tracked on the right. The y-axis uses a linear scale, and the height of the bar reflects the fraction of each class of mutations that was not eliminated by selection. A value of 1.0 is the expected height for all bars when there is no selection. Perfect selection for a given mutation effect interval will result in bars of zero height.

Figure 2b shows the beneficial half of this same distribution. The distribution extends from the left with the smallest mutational effects tracked (tracking threshold) to the maximal beneficial effect (on the right), and any bar significantly higher than 1.0 reflects positive selection. Note: These figures are not generated until there are a “sufficient” number of mutations. Even so, plots can initially show significant noise associated with sampling error until larger data sets have built up. This is especially true for beneficial mutations (Figure 2b), because they are usually relatively rare. The user can know that

the number of accumulated mutations has become large enough to give reliable plots when there is little random fluctuation in bar heights and the transition from selectable to near-neutral becomes smooth and unambiguous.

Figure 3: This figure plots the selection threshold for dominant alleles as a function of the generation count, beginning with generation 200. The selection threshold, described more completely in the glossary below, is the value of absolute fitness effect for which the deleterious allele frequency is 50% of what it would be, if there had been no selection. For values of absolute fitness effect smaller than this threshold, alleles frequencies are governed more by drift and less by selection, while for values larger than the threshold, the opposite is true. Data for recessive alleles is not plotted, but is available in the data file. At any given generation, one can readily estimate the selection threshold visually using Figure 2, by observing the fitness value on the horizontal scale corresponding to a bar height of 0.5.

Figure 4: This figure, like figure 2, is also designed to reveal how selection is altering the distribution of mutation effects in the population. By using a linear (but truncated) scale for the x-axis, it shows more closely just the lower impact mutations. It displays, in greater detail the differences between the theoretical (red) distribution of mutation effects (as would accumulate apart from selection) and the actual distribution of accumulating mutations (blue = recessive, green = dominant). Like figure 2, this figure reveals the variable effectiveness of selection over a range of mutation effects. Unlike Figure 2, the viewer can readily see that small-effect mutations are always much more numerous than large-effect ones, both before (red), and after (blue/green) selection. Figure 4a displays deleterious mutations, and figure 4b displays beneficial mutations. Like figure 2, the plotting of accumulated mutations is only reliable when enough mutations are present, evident when the distribution visually transitions from jagged to smooth.

Figure 5: This figure displays the frequency distribution of the fitnesses of the accumulating haplotypes versus their composite fitness effect. Those haplotypes with a net deleterious effect are plotted in red (left of zero) and those with a net beneficial effect in green (right of zero). This figure shows the distribution of net linkage blocks fitness effects (as opposed to individual mutation effects).

Figure 6: This figure superimposes the distribution of phenotypic fitness values of the individuals in the population before (red), and after (green), selection. A histogram format is used. The blue line plots the ratio of the number of surviving (selected) individuals within a given phenotypic class versus the number of offspring in that fitness category prior to selection (scale shown on the right). The portion of the red distribution not covered by the green represents those offspring that are selected away (those offspring that will not reproduce to create the next generation).

Figure 7: This figure displays the frequencies of all tracked individual mutant alleles in the population. It is plotted at generation 200, at generation 1000, then every 1000 generations, and finally at the end of the run. The figure provides the number of very rare alleles, in the population (having frequencies of less than 1%), the number of

polymorphic alleles (having frequencies between 1- 99%), and the number of fixed alleles (having frequencies greater than 99%). The data button on this plot provides access to the data for the current and all the previous plots. The file with suffix .pmd provides even more detailed information on the distribution of polymorphisms at each of these plot times. Results are in table format with 500 values of polymorphism frequency (50 intervals) vs. fitness effect (10 intervals). Caution – this data represents only tracked mutations, so the tracking threshold must be set to be less than the lowest-impact mutation effect for this data to be fully accurate.

Applications of MENDEL.

Teaching: The MENDEL program is a useful teaching tool to demonstrate to students in a concrete and visual manner the fate of mutations once they enter a population, and how they increase in frequency, are eliminated, or simply drift randomly. MENDEL shows how these dynamics play out over many generations under a wide range of conditions. The student can see how this process affects average mutation count per individual, average fitness, allelic frequencies, and mutational fixations over time. The student can experiment with the biological parameters that alter the rates of these processes. MENDEL also allows the student to see exactly what happens during a population bottleneck, what happens in a mutational meltdown scenario, how genes can circulate between sub-populations, and what happens when key biological parameters are modified during a run.

Research: MENDEL's Accountant can function as a sophisticated research tool. To our knowledge, there is no simulation program comparable that provides genetic researchers with such a realistic and flexible research simulation capability. Highly specific scenarios can be run that have bearing on extinction of species, management of endangered species, germplasm preservation, epidemiology, ecology, etc. Likewise, simulations can also be run which have bearing on more basic questions, including the relative importance of the different variables that affect selection efficiency.

MENDEL Glossary.

1. Mendel's Accountant (MENDEL). An advanced numerical simulation program that acts as a genetic accounting system, and realistically models how genomes change over time in response to mutation and selection.
2. Genome. The entire genetic content of an organism. In MENDEL, the initial genome is not specified except in terms of genome size, chromosome number, and number of linkage blocks. A functional genome is simply assumed as a backdrop for the accumulating mutations, which are individually being tracked.

3. Functional genome. The entire physical genome, minus any portions that have no biological expression or consequences relating to biological fitness.
4. Population. All the individuals that constitute an inter-breeding group. In MENDEL, the population is specified in terms of number of reproducing individuals, mating pattern, fertility level, and sub-structure (tribes).
5. Mutation. Any heritable change in the genome that was not present in the previous generation. In MENDEL, mutations can have a range of effect from lethal to beneficial, and a range of expression from entirely dominant to entirely recessive. MENDEL tracks the effect of every mutation from the time that the mutation enters the population until it may be lost due to selection or random drift.
6. Mutant locus - The location of a mutation, in terms of its position within a linkage block within a chromosome. In MENDEL, all loci are assumed to be non-mutant except where a mutation has been added. Once a single mutation arises within an individual at a specific location, the same corresponding location in all the other individuals not carrying this mutation, by definition, becomes the non-mutant allele.
7. Mutant allele. All the derived copies of an initial mutation, which are being passed from generation to generation. A mutant allele can increase or decrease in its frequency within the population.
8. Mutant allele frequency - The mutant allele frequency for a given locus is determined by number of copies of a given mutation in the population, compared to how many copies there would be if every individual was homozygous for that mutation. (For diploids the total number of possible mutant copies is two times the population size.). If there are 2 copies of a mutation in a diploid population of 100, the mutant allele frequency is 1%, so the non-mutant allele frequency is 99%.
9. Mutation fitness effect. We refer to the biological impact of a mutation on individual fitness as the mutation's "fitness effect". In MENDEL a given mutation fitness effect can be small or large. The mutation effect is expressed as the relative change in an individual's total *biological functionality*, as reflected by a corresponding change in an individual's *genotype value* (see below). A deleterious mutation with an effect of -0.01 decreases an individual's genotype value by 1%. Crudely speaking, one percent of the genomic information is lost, or more accurately, total biological functionality is reduced by 1%. Another way of saying this is that such a mutation decreases genotypic fitness by 1%. These are all just alternative ways of describing the biological effect of a mutation. Mutation effect is independent of environmental variation (phenotypic noise), and random aspects of reproduction (reproductive noise). Mutation effect is similar but not identical to the traditional concept of a *selection coefficient*. See Sanford et al. (SCPE 8(2), 147-165) for the mathematical function we use to generate the distribution of mutation effects, and for an explanation of the difference between fitness effect and selection coefficient.
10. Genotype. The genotype is the specific collection of genetic alleles present in a specific individual within the population. In MENDEL, the starting genotype for all individuals is an unspecified and invariant genome for the organism. MENDEL specifies only the mutational deviations from this non-mutant starting

genotype. In MENDEL, the specified genotype is simply the sum total of all the mutant alleles (including their chromosomal locations), within an individual.

11. Genotype value. The genotype value is that portion of an individual's total biological functionality that is derived exclusively from that individual's genetic makeup. The genotype value is different from the phenotype value because environmental factors (phenotypic noise) also contribute to an individual's biological functionality.

    In MENDEL, the initial genotypic value of all individuals is defined as 1.0. Beneficial mutations increase this value, and deleterious mutations decrease this value. The extent to which a mutation alters genotypic value is a function of its specific mutation effect (see above). Genotype value can be understood as being synonymous with the term *genotype fitness*. However, the concept of genotype value (genetic fitness) is distinct from what most population geneticists formally define as "fitness". For clarity, we will use the term *reproductive fitness* (see below), to refer to the traditional population geneticist's definition of fitness, as distinct from *genetic fitness*. Genetic fitness is what is actually plotted in MENDEL's figure 1b.

12. Linkage block (haplotype). Mutations are not inherited independently but are passed from generation to generation in clusters or blocks. These clusters are physically linked together, within "linkage blocks", which represent specific regions of linear chromosomes. Although chromosomes recombine something like cutting a deck of cards, some cards consistently stick together, due to recombinational "cold" spots. Points of frequent recombination (hot spots) separate linkage blocks (cold spots) from each other. A specific set of mutations which is physically being inherited as a single unit is called a *haplotype*.
13. Phenotype. The actual biological functionality of an individual. The phenotype is affected both by the genotype and by the environment in which the individual develops. The genotype and phenotype are correlated, but they are not identical. In MENDEL, the phenotype value (or fitness) is created by adding to each genotype value (or fitness) a random "environmental noise value" based upon the specified "heritability" and using a random number generator. This environmental noise value is not heritable and so is not passed on to the next generation.
14. Phenotype value. This is the actual biological functionality of an individual, arising due to the combination of genotypic effects and environmental effects. Phenotypic value is what selection actually acts on - it is what "Mother Nature" actually "sees". In MENDEL, the initial *mean* phenotypic value is always 1.0. In the initial first generation, all individuals will have an identical genotype, but there will still be variance around the population's mean phenotypic value - due to environmental effects. The amount of environmental noise is controlled by specifying a "heritability value", which is the input parameter that defines the ratio of genotypic variance to environmental variance. Because the phenotypic contribution from heritability scales with genotype value and therefore becomes small when the genotypic value becomes small, an additional non-scaling noise factor can also be used to modify phenotypic values

Phenotype value is synonymous with the terms *phenotypic fitness* or *biological fitness* as reflected by the common use of these terms among biologists. However, the concept of phenotype value (phenotypic fitness) is distinct from what population geneticists formally define as "fitness". For clarity, we will use the term *reproductive fitness* (see below), to refer to the traditional population geneticist's definition of fitness, which is distinct from phenotypic fitness.

15. Reproductive fitness. We define this term as the phenotypic value (phenotypic fitness), plus "reproductive noise". Reproductive noise arises because actual success in reproduction is not just determined by biological functionality, but also by random reproductive factors. So phenotypic fitness and reproductive fitness are correlated, but not identical. The strength of correlation between phenotype value and reproductive fitness depends upon the selection scheme employed. Artificial truncation will yield the highest possible correlation, while classical probability selection will yield the lowest correlation of the various schemes implemented in MENDEL. What we are calling "reproductive fitness" is sometimes called "Darwinian fitness" or "Wrightian fitness" after Sewell Wright, the first to formulate "Darwinian fitness" mathematically. We recognize that "Darwinian fitness" encompasses elements beyond "reproductive fitness", but have not identified a more precise term that is still understood intuitively by a broad audience.
16. Selection threshold. The selection process eliminates deleterious mutations with large negative fitness effect values more effectively that it does for deleterious mutations with smaller ones. Similarly, selection enhances the frequencies of favorable mutations with large positive fitness effect values more effectively than it does for favorable mutations with smaller values. For fitness effect values sufficiently small, selection plays essentially no role in altering the frequencies of such alleles. Mutations in this range, both deleterious and favorable, are generally referred to as "effectively neutral". The fate of these mutations is governed essentially entirely by drift. By contrast, deleterious mutations that have large impacts on genetic fitness are eliminated very effectively by selection such that their frequencies in the population are maintained at nearly zero. Typically, there is a very broad transitional zone (representing several orders of magnitude of fitness effect), between the zone of highly effective selection and the zone of essentially no selection. The mutations in this transition zone have been termed "nearly neutral". We define the term 'selection threshold' as the absolute value of fitness effect at the midpoint of this transition region - wherein the allele frequency is precisely 50% of what it would be if there had been no selection. This means that above this threshold value in absolute fitness effect there is more than 50% elimination of deleterious mutations as a result of the selection process, while below this threshold value there is less than 50% elimination. Typically, mutations more than an order of magnitude smaller in absolute fitness effect below the threshold are not significantly influenced by selection, while deleterious alleles more than an order of magnitude larger than the threshold are entirely eliminated by selection. For the case of favorable alleles, the same absolute threshold value tends to apply.

17. Mutation/selection chain. In the real biological world, this is the chain of events that links a mutational event to a selection event. A single mutation affects a linkage block, which affects a chromosome, which affects a genotype, which affects a phenotype, which affects the reproductive fitness of an individual, which affects the actual transmission of a mutation into the next generation. There is biological noise at each link in this chain, and so each link of this chain is associated with an imperfect correlation. MENDEL is designed to accurately reflect this chain of events, so that a given mutant value actually has a limited effect on a linkage value, which has a limited effect on a chromosome value, which has a limited effect on a genotype value, which has a limited effect on a phenotype value, which has a limited effect on the reproductive fitness value of a given individual - which defines the actual transmission of the mutation. The strength of the correlation at each stage depends on the values chosen by the user for the various input parameters.
