# Computer Simulations to Model Mendel's Laws on Inheritance in Computational Biology

Gurmukh Singh, Ph.D.
Department of Computer and Information Sciences
SUNY at Fredonia, Fredonia, NY 14063 USA
singh@cs.fredonia.edu

## ABSTRACT

The present paper uses the computational power of Microsoft Excel 2007 software system to perform interactive simulations for advanced undergraduate and first year graduate students in biology, biochemistry, biophysics, microbiology and medicine in college and university classroom setting. We developed the required computer algorithm, which uses a built-in pseudo-random number generating function, to computer model two basic Mendel's Laws of heredity in evolution for plant and animal species. The results of simulations corroborate the experimental observations of Mendel's research on inheritance in *Pisum hybrid species*.

## Categories and Subject Descriptors

I.5.4 [**Pattern recognition**], H.4.1 [**Microsoft Excel**]: Applications – *Mendel's laws, Bioinformatics*

## General Terms and Keywords

Algorithm, Theory, Mendel's Laws, Monte Carlo Simulations, Recessive and Dominant Genes, Genetics; MS Excel 2007

## 1. INTRODUCTION

Mendel's famous research paper [1] on inheritance in *Pisum* hybrids formed a sound basis for the study of modern genetics, but the importance of Mendel's work was not completely realized till the beginning of twentieth century. In 1851, he was sent by the Church to study natural science at the University of Vienna, so that he could defend the theory of intelligent design. After completing his studies, he systematically and painstakingly investigated for seven years 28,000 species of *Pisum* hybrids in his father's orchard with an idea of developing new varieties of plant and sheep through cross-breeding. In 1865, he presented his research findings to the Natural History Society of Brünn, and the following year, he published his most prominent paper on the laws of dominance, segregation, and independent assortment [1]. Mendel himself was very much interested in the question of evolution theory of Darwin [2], but unfortunately, he performed experiments with an idea in his mind to support the theory of intelligent design. He worked in the tradition of Kölreuter and Gärtner [3], investigating Linnaeus's theory [4] that hybrids might have played an important role in evolution process and consequently, his experiments were designed to explore a fundamental difference between hybrids and species. Mendel believed that the mechanisms of evolution might have existed in the

organisms themselves, but they could not be directly observed, so it was necessary to investigate the aggregate behavior of a large amount of data collected and to analyze these data using statistical techniques, concluding that mathematics should play an important role in the formulism corresponding to any experimental scientific research.

The basic aim of Mendel's experiments on hybrids was to provide the detailed evidence that his predecessors were unable to figure out. To achieve his goal, he initiated discussion of the evolutionary significance of hybrids by making a clear distinction between "variable" and "constant" hybrids [5]. The progeny of variable hybrids would display reversion, while constant hybrids would breed exactly the same way. Mendel's ultimate objective was to find out which environment could give rise to which types of hybrids, and whether or not constant hybrids could be produced in a reproducible experiment. Mendel's research was largely ignored by others, although it was not completely unknown to biologists of the time. Mendel thought that his laws could be applied only to certain categories of special species such as *Pisum* hybrids and not to other general classes of species. Nevertheless, controversial nature of Mendel's laws was vigorously promoted in Europe by William Bateson [6], who coined a new term "genetics", "gene", and "allele" to describe many of its traits. The model of heredity was highly debated by other biologists as it implied that heredity was discontinuous, contrary to the apparently continuously changing observables. Several biologists also did not care about Mendel's theory for numerous years, since they were not sure that it could be applied to all plant and animal species. Consequently, they generally concluded that there seemed to be very few true Mendelian characters existing in nature. However, later research done by biologists and statisticians like Fisher [7] gave some concrete evidence that if multiple Mendelian factors were involved for individual traits, they could produce the diversified results observed in nature overtime. Morgan [8] and his associates later on integrated the theoretical model of Mendel with the chromosome theory of inheritance, where it was assumed that the chromosomes of cells were thought to hold the actual hereditary information, and create what is now popularly known as classical genetics [9]. Thus, classical genetics of Morgan [9] was extremely successful in establishing Mendel's laws of heredity on a firm footing.

As pointed out by Fisher [7] if the results of Mendel's experiments on *Pisum* hybrids [1] are carefully examined, one would immediately conclude that it was highly desirable to formulate a mathematical model, which would reproduce the experimental results on hybrids. Although mathematics, computer science and

statistics are not considered to be a part of natural sciences, strictly speaking, computer science (in conjunction with modern mathematical computational tools such MS Excel 2007, MatLab, Mathematica, Maple etc.) provides essential algorithm and logic in terms of high level programming languages such as C++, C#, Visual BASIC (VB), Java, Pearl etc. to construct successful and relevant model of any experimental scientific research or discovery. Thus, mathematics, statistics and computer science provide basic essential and viable tools and framework to model theoretically or empirically almost all the natural phenomena of modern scientific exploration, including two Mendel's laws of inheritance. This is the ultimate aim of the present research, in which we intend to use a MS Excel 2007 to model Mendel's laws of heredity in genetics.

Although a lot of work has been done in literature to build theoretical models to explain Mendel's laws of heredity [10], very scanty work has been done using Microsoft Excel 2007. As far as we are aware of, only one investigator has employed Excel 2003 to build model on Mendel's laws of inheritance, which explains the behavior of recessive genes only [11] and none of the investigators, so far, have tried to build theoretical and empirical model on the behavior of dominant genes to verify Mendel's laws of genetics.

The paper is organized as follows: Section 2 is devoted to a brief description of MS Excel 2007 computational and graphical capabilities. Section 3 deals with the theoretical formalism and algorithm for Mendel's laws of heredity, and it also reports actual computer simulation work on recessive and dominant genes in tabular and graphical form. Finally, conclusions of the present investigation are presented in Section 4.

## 2. MS EXCEL 2007 SPREADSHEET AND ITS "INSERT FUNCTION" GROUP

Before starting the formal computer simulation work on interactive application of MS Excel 2007 to model Mendel's laws of heredity, a few words about Excel's 2007 spreadsheet, and its "*Insert Function*" Group may be in order. Each spreadsheet in Excel is composed of 256 columns, 65536 rows and one may insert a large number of spreadsheets in a workbook [12], the number of rows in a single spreadsheet is a measure of the number simulations that can be performed during a single computation, i.e., almost 66000, which was not possible even with old versions of IBM Mainframes of seventies.

Figure 1 shows an actual screen-shot of MS Excel 2007 "*Insert Function*" Group, which is very natural and the students/instructors can grasp its mastery pretty easily. Figure 1 also displays one category of selected functions, e.g., "*Logical*" category, which is highlighted in gray color. To its left side, there are different kinds of functions to be selected from "*Select a function*" window. One can see all the built-in Excel 2007 functions corresponding to "*Logical*" function category and the AND( ) function is ready for use. For the present investigation two functions were employed from the "*Logical*" category: IF( ), and AND( ) functions. In addition, three more built-in functions are used in this study and these are: COUNT( ), SUM( ) and RAND( ) functions. For further details of MS Excel interfaces and graphic capabilities, reference may be made to our former investigation [12].

## 3. MODELING OF MENDEL'S LAWS AND DISCUSSION OF RESULTS

To perform the simulations to produce off-spring or progeny through the fertilization of sex chromosomes of males ($\male$) and females ($\female$), we start with four pairs of dominant and recessive genes, and we use the idea of Mendel's Laws for *Pisum* hybrids [1]. If we begin with two pairs of genes DR, where, D and R stand for dominant and recessive genes, respectively. In two pairs of genes, there are four possible combinations: DD, DR, RD and RR or DD, 2DR, RR. As only two genes are involved in the fertilization process of an egg with a sperm, each gene has 50% contribution for the production of progeny and this condition must be implemented in the computer simulations. The RR kind of progeny has less likelihood of survival in the subsequent generation, and hence the ratio of their population in successive generations should stay constant as observed by Mendel through his research on *Pisum* hybrids [1]. This is what is found through the computer simulations of data presented in the last column of Table 1, where this ratio has constant value equals 2. The ratio is also plotted in Figure 2 for nine generations and shown with filled triangles (▲). The plot shown with filled squares (■) seems to deviate from the one for the participating genes for producing the off-spring, indicating less growth in future generations and will not survive overtime.
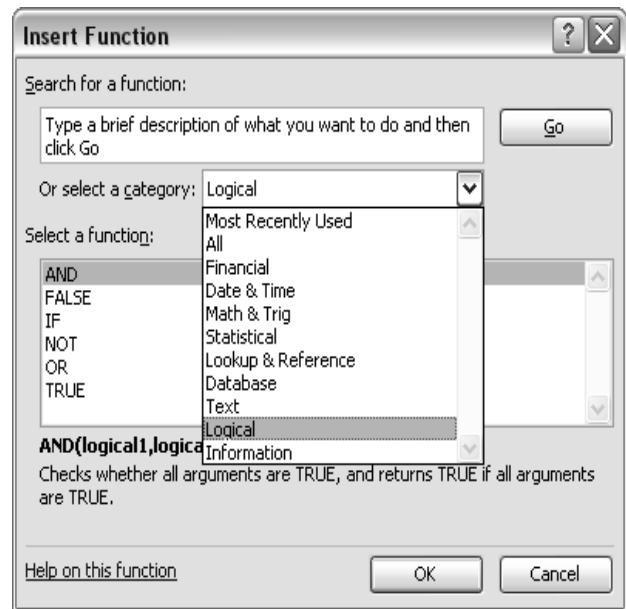


**Figure 1: A typical view of MS Excel 2007 of "*Insert Function*" Group, in which various built-in logical functions can be used during computational work.**

The best fit exponential function through the data points (filled diamonds) for the sum total number of simulations for each generation of recessive genes is given by

$$y = C_1 \, e^{0.696x}, \qquad (1)$$

where $C_1 = 7.585$ and $e_1 = 0.696$. Exactly the same way, the first best-fit exponential function through the data points for the total number of progenies (filled squares) for each generation of recessive genes is given by

**Table 1: Simulations of the contribution of recessive genes in the population growth for several generations are displayed in Table 1, where ratio of the progeny growth in two successive generations stays constant, which is in accordance with Mendel's Laws [1]. This table shows only 11 simulations out of a total of more than 16,000 simulations. In column four, 1 stands for the progeny due to recessive gene-pair, whereas 0 stands for the off-spring due to dominant gene-pair.**

| Female Gene | Male Gene | Progeny Gene | Recessive/Dominant | Recessive Genes | | | |
|---|---|---|---|---|---|---|---|
| dominant | dominant | dominant | 0 | Generation | Trials | Progeny | Ratio |
| recessive | recessive | recessive | 1 | 1 | 15 | 4 | 2 |
| recessive | dominant | dominant | 0 | 2 | 30 | 9 | 2 |
| dominant | recessive | dominant | 0 | 3 | 62 | 17 | 2 |
| dominant | recessive | dominant | 0 | 4 | 125 | 34 | 2 |
| recessive | recessive | recessive | 1 | 5 | 250 | 63 | 2 |
| recessive | dominant | dominant | 0 | 6 | 500 | 123 | 2 |
| dominant | recessive | dominant | 0 | 7 | 1000 | 245 | 2 |
| recessive | dominant | dominant | 0 | 8 | 2000 | 508 | 2 |
| recessive | recessive | recessive | 1 | 9 | 4000 | 1014 | 2 |
| recessive | dominant | dominant | 0 | 10 | 8000 | 2042 | 2 |
| recessive | recessive | recessive | 1 | 11 | 16000 | 4089 | |

$$y = C_2\, e^{0.695x}, \qquad (2)$$

where the constant, $C_2 = 1.937$ and exponent, $e_2 = 0.695$. It will be a good idea to compute the ratio of constants $C_1$ and $C_2$. This ratio $C_2/C_1 \approx 25\%$, which proves that only a quarter of the total population in a given generation comes from the contribution of recessive genes, and rest of the population in the same generation is from the contribution of dominant genes, which is pretty interesting outcome of the present simulations. The two exponents, $e_1$ and $e_2$, in Eq. (1) and Eq, (2) have exactly the same value as expected, which indicates that the population of recessive gene pairs will not increase or explode overtime. Thus, the progeny due to participation of recessive gene pairs will not survive overtime too. This is what is the conclusion obtained from Darwin's theory of evolution [2].

As shown by Mendel that the population growth of progeny produced by the contribution of dominant genes in a given generation, $P_n(D)$, should be expressed in terms of the progeny production by the recessive genes, $P_n(R)$. Following similar arguments of Ref. [1], one can write an empirical mathematical correlation between the two progenies in the first generation with $n = 1$:

$$P_1(D) = \Sigma\, P(\text{Dominant}) - \Sigma\, P(\text{Recessive}). \qquad (3)$$

Here, $\Sigma\, P(\text{Dominant})$ and $\Sigma\, P(\text{Recessive})$ stand for the sum total population in the first generation as a result of the fertilization of eggs and the sperm cells of dominant and recessive genes of their parents.

For second and subsequent generations, the correlation between the population growth due to dominant and recessive gene progenies can be expressed as

$$P_n(D) = K(2^n - 1)P_n(R), \qquad (4)$$

where constant $K = 4$ and $n = 2, 3, 4, 5, 6, \ldots\ldots$etc. is a running

index for $2^{nd}$, $3^{rd}$, $4^{th}$, $5^{th}$, $6^{th}$, …etc. generation, respectively.
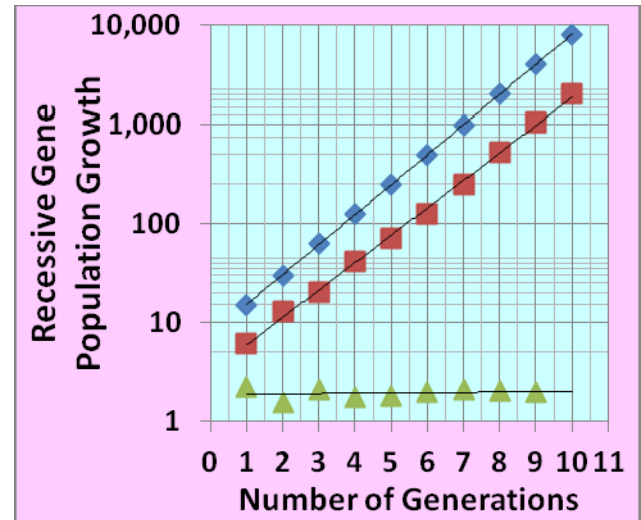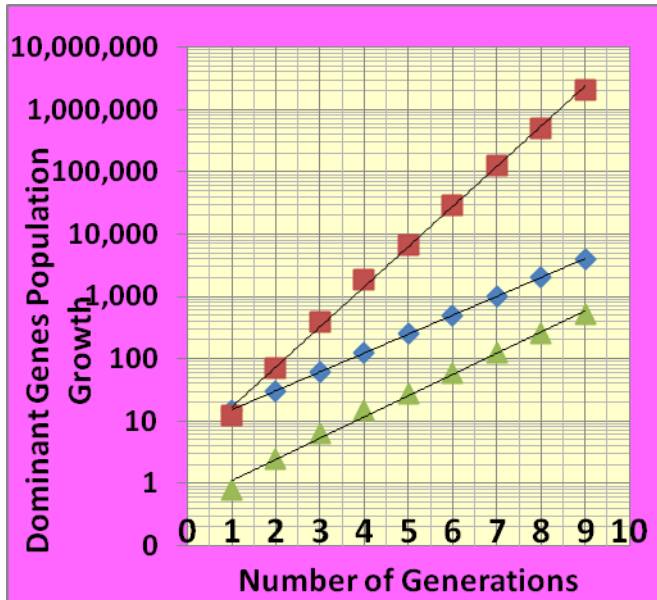


**Figure 2: A plot of population growth of progenies due to recessive genes and ratio of baby population in successive generations as a function of the number of generations. Symbols used in this diagram are: (1) diamonds (♦) represent starting dominant/recessive gene pairs for a given generation, (2) squares (■) are for population growth due to recessive genes and (3) triangle (▲) is the ratio of two successive progenies due to recessive genes.**

Multiplier factor, $(2^n - 1)$ used in expression of Eq. (4) has odd values, which are shown in $2^{nd}$ column of Table 2. The results of these computations are displayed in Table 2 and are plotted graphically in Figure 3. It is clear from Table 2 that starting from the second generation and for subsequent generations, the ratio of off- spring in a given generation to the number of

**Table 2: Computer Simulations for the contribution of dominant genes in the population growth in several generations are displayed in Table 2, where ratio of the progeny population growth increases in accordance with Mendel's Laws [1].**

| Dominant Genes (Constants) | | | | |
|---|---|---|---|---|
| Generation | Multiplier | Staring Dominant/Recessive gene pairs, $P_n(DR)$ | Progeny, $P_n(D)$ | Ratio, $P_n(D)/P_n(DR)$ |
| 1 | 1 | 15 | 12 | 0.8 |
| 2 | 3 | 30 | 72 | 2.4 |
| 3 | 7 | 62 | 392 | 6.3 |
| 4 | 15 | 125 | 1860 | 14.9 |
| 5 | 31 | 250 | 6696 | 26.8 |
| 6 | 63 | 500 | 29484 | 59.0 |
| 7 | 127 | 1000 | 125984 | 126.0 |
| 8 | 255 | 2000 | 495720 | 247.9 |
| 9 | 511 | 4000 | 2023560 | 505.9 |
| 10 | 1023 | 8000 | 8081700 | 1010.2 |
| 11 | 2047 | 16000 | 32473608 | 2029.6 |

simulations in the same generation increases exponentially, which indicates that the population due to dominant genes should increase or explode in subsequent generations, as observed by Mendel in *Pisum* species. The population in the second generation is six times than that in the first generation. The ratio of populations in third and second generations is almost four times, and this is also true for the ratio of almost all the subsequent generations. These findings corroborate the experimental results of Mendel on *Pisum* hybrids [1].



**Figure 3: A plot of population growth of progenies due to dominant genes and ratio of progeny population in successive generations as a function of number of generations. Symbols used in this diagram are: (1) diamonds (♦) represent starting dominant/recessive gene pairs for a given generation, (2) squares (■) are for population growth due to dominant genes and (3) triangles are (▲) the ratio of progeny due to dominant genes to the starting pairs of dominant/recessive genes.**

The best-fit exponential functions through the three sets of data points are represented by Eq. (5) – Eq. (7). Each function when plotted on semi-log scale looks like a straight line plot as depicted in Figure 3. The first best-fit exponential function that passes through the dominant/recessive gene-pair data points (♦) for all the generations is given below:

$$y = C_1' e^{0.696x} \qquad (5)$$

Here, once again, $C_1' = C_1 = 7.585$ and exponent has a value $e_1' = 0.696$, which is exactly the same as in Eq. (1). This function also has exactly the similar form as that of Eq. (1), because Eq. (1) and Eq. (5) represent the same number of starting dominant/recessive gene pairs for a particular generation.

The second best-fit exponential function passes through the dominant gene pairs (■) for all the generations and is given below:

$$y = C_2' e^{1.482x} \qquad (6)$$

Here, $C_2' = 3.801$ and $e_2' = 1.482$. By comparing constants $C_2'$ and $C_1'$ through their ratio $C_1'/C_2' = 1.996$, we obtain an important information of the intercept of each plot, i.e., the intercept of dominant gene progeny population graph is two times more than the starting dominant/recessive gene-pairs plot. In the same way, the ratio of exponents $e_2'/e_1' = 1.482/0.696 = 2.129$ indicates that the slope of dominant gene progeny population plot is a little more than the starting dominant/recessive gene-pairs plot, which is also obvious from Figure 3.

The third best-fit exponential function in Figure 3 is for the ratio of progeny corresponding to the dominant gene pairs, $P_n(D)$, to the starting population of dominant/recessive gene-pairs,

$P_n(DR)$:

$$y = C_3^{'} e^{0.784x}, \tag{7}$$

where $C_3^{'} = 0.504$ and the exponent in Eq. (7) is $e_3^{'} = 0.784$. The behavior of ratio of progeny of the dominant gene pairs, $P_n(D)$, to the starting population of dominant/recessive gene-pairs, $P_n(DR)$ has very similar nature as that of the starting population of dominant/recessive gene-pairs, $P_n(DR)$. This conclusion can be easily drawn from Figure 3, in which two plots represented by the data points corresponding to diamonds (♦) and triangles (▲) are almost parallel to each other, showing that their slopes should be almost equal. If one looks at the values of $e_1^{'} = 0.696$ and $e_3^{'} = 0.784$, one can perceive that their values are almost match and therefore their slopes are almost equal.

## 4. CONCLUDING REMARKS

In the current investigation with the latest version of MS Excel 2007, we have successfully presented for advanced undergraduate and first year graduate classroom setting, the interactive simulations of Mendel's laws of heredity. We start with two pairs of dominant and recessive genes, perform the computer simulations with help of pseudo-random number generating function, and performed more than 16,000 Monte Carlo simulations. We implemented the condition of 50% probability of fertilization from each dominant and recessive gene-pair. Simulations were performed for more than ten generations of population of recessive and dominant genes. The results of the simulations for recessive gene progeny corroborate the predictions of Mendel's laws on *Pisum* hybrids. Theoretical formalism has been extended to create empirical recursive formula for the progeny of dominant genes, and the theoretical results, once again, match with predictions of Mendel's laws of inheritance. Both for recessive and dominant gene progenies, graphs of their population growth are plotted as a function of the number of generations. The least squares, best-fit exponential functions have been obtained for recessive and dominant gene progenies and their biological implications have been discussed to support the implosion of population of progeny from the recessive gene parents, and explosion of progeny from the dominant gene parents. The present investigation proves that progeny with recessive genes will not survive in successive generations. However, the progeny from the dominant genes parents shall survive and it will explode overtime. These findings are quite in agreement with Darwin's theory of survival of the fittest [2].

Although the present investigation was undertaken from the perspective of advanced undergraduate and first year graduate students, we hope that it might have important implications for biosciences, natural science, computer science and medical science educators as well: (i) In biological and medical sciences the instructors can demonstrate in class-room setting how to employ versatile software system such as the latest version of MS Excel 2007 [12] to simulate the basic concept of Mendel's laws of heredity, whereas (ii) in computer science, and natural science instructors as well students of these disciplines could visualize the real time application of this fundamental concept of botany, biology and other medical sciences in a virtual laboratory. In future, we would like to test Mendel's laws experimentally on species like *drosophila melanogaster*.

## 6. REFERENCES

[1] G. Mendel, "Experiments on Plant Hybrids." In: The Origin of Genetics: A Mendel Source Book, (1866).

[2] C. Darwin, "On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life," p. 162 (1859)

[3] J. G. Kolreuter, In *Encyclopædia Britannica* (2008). Retrieved March 2, 2008, from Encyclopædia Britannica Online: http://www.britannica.com/eb/article-9045956

[4] J. L. Larson, **"**Linnaeus and the Natural Method**"**, *Isis*, Vol. 58, No. 3, pp. 304-320 (Autumn 1967).

[5] L. A. Callender, Gregor Mendel: An opponent of Descent with Modification. *History of Science* 26, 41-75 (1988).

[6] W. Bateson, Materials for the study of variation: treated with special regard to discontinuity in the origin of species, Macmillan and Co., London (1894).

[7] http://www-history.mcs.st-andrews.ac.uk/ /Extras/Fisher_Statistical_Methods.html

[8] G. Allen, Thomas Hunt Morgan, pp. 1-447, Princeton University Press, Princeton, N.J. (1978); Sturtevant, A.H., A History of Genetics, pp. 1-165, Harper and Rowe, New York (1965).

[9] Classical genetics consists of the techniques and methodologies of genetics that predate the advent of molecular biology. A key discovery of classical genetics in eukaryotes was genetic linkage.

[10] L. G. Robbins, Do-It-Yourself Statistics: A Computer-Assisted Likelihood Approach to Analysis of Data from Genetic Crosses, Genetics, Vol. 154, 13-26, January 2000; F. Schoonjans, A. Zalata, C. E. Depuydt, and F. H. Comhaire, MedCalc: a new computer program for medical statistics, Comput. Methods Programs Biomed. 48(3):257-62 (1995); G. R. Margarid, A. P. Souza and A. A. Garcia; OneMap: software for genetic mapping in outcrossing species, Hereditas, 144(3):78-79 (2007); P. A. Nuin, WinPop 2.5: software for representing population genetics phenomena, Brief Bioinform, 6(4):390-3 (2005).

[11] Introduction to Interdisciplinary Computational Science Education for Educators, SC07 Education Program Summer Workshop Series, Buffalo State College, Buffalo, NY, June 3-9, 2007.

[12] Microsoft Excel Software Usage for Teaching Science and Engineering Curriculum, G. Singh and K. Siddiqui, Journal of Educational Technology Systems, Vol. 34 (4), 405-417 (2009).