BACKCROSSING: A MATHEMATICAL ANALYSIS OF GENE INSERTION IN EXISTING HYBRIDS AND STATISTICAL VALIDATION

TERRENCE P MCGARTY

Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Cambridge, MA

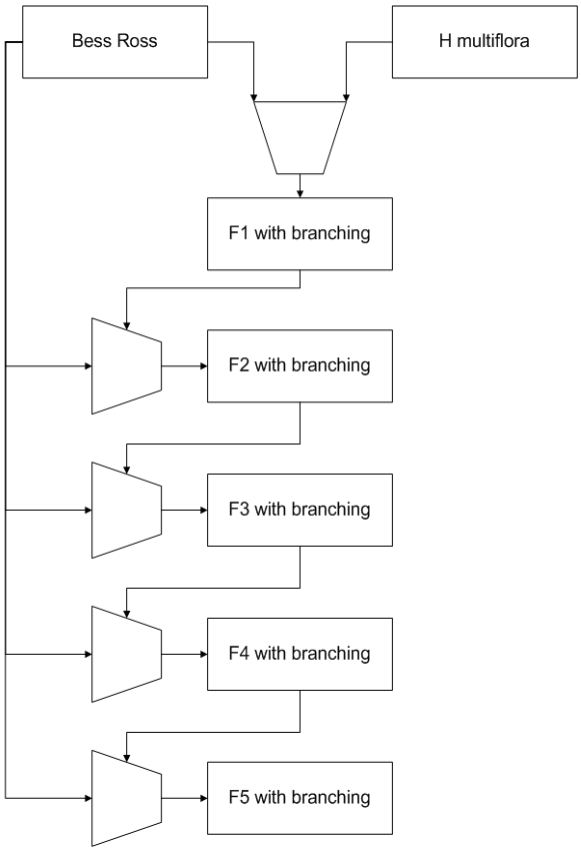
Backcrossing has been used for centuries. It is however frequently misunderstood and misapplied. In addition there appears to be limited mathematical models for the process of backcrossing and there thus results limited understanding of its application and capabilities. In this paper we review backcrossing using a specific Genus, Hemerocallis, and ten we develop a detailed mathematical model to analyze backcrossing in a generalized format. One of the key issues to be addressed is that of how many generations are required to assure an effective backcross, namely insertion of a desired gene, and the corollary question of how well this can be determined by a statistical analysis of the resulting backcrossed offspring. We also examining the inverse problem of estimating the number of operative genes which control the phenotypes based upon the measured results. Along with this problem we develop bounds on the accuracy of the estimation procedures.

Backcrossing is a simple process. One takes a plant with characteristics one is comfortable with, and then seeks to introduce a new characteristic from some other plant into the original one. For example, we may take the hybrid "Bess Ross", a diploid red daylily and seek to introduce into the plant a branching as one may find in the species H multiflora. We desire only the branching characteristic of H multiflora and we desire to retain all other characteristics of Bess Ross. The process we would employ would be backcrossing.

Backcrossing then works as follows. We first select a plant whose features we are satisfied with but for one characteristic. In our example we start with a diploid hybrid named Bess Ross, a red flower with no substantial branching. We want to introduce extensive branching into the plant. We want just the branching and not any of the other characteristics. Thus we say we desire to "drive" or insert the single characteristic of branching into the target plant. After the first cross, we then cross selected offspring, namely those with branching, with Bess Ross, again and again. After M such crosses we then ask what is the probability that we have the desired branched but otherwise homozygous Bess Ross. The result is then a plant which we could reproduce from seed and have a high level of confidence that it will breed true to form; namely a branched red flower appearing as a Bess Ross.

There has been an extensive amount written on backcrossing. The classic work of Allard uses a simplified two gene model and tries to exemplify the process. We argue herein that one must deal with the complex multi-gene model and no just two genes. The important issues result only when considering N genes. The recent work of Brown and Caligari also address the issue the same way. The results are frankly deceptive at best. The use of the approach in hybridizing horticultural plants requires a broader understanding of the issues. The work of Mayo also attempts to summarize the literature but we feel it too falls quit short of what is required. Brown et al also examine the issue but again do not address the details of the statistical model or the generalizations required. Similar high level analyses are performed by Griffiths et al as well as by Strickberger but failing in detail and depth.

The flow chart below depicts the details of standard backcrossing. It will be this process which we will analyze in some detail.



MATHEMATICAL MODELS

We start with the Recurrent plant, in this case the "Bess Ross" red diploid. It is assumed to have a collection of genes which control the flowering mechanism; These genes are assumed to control color, branching, budding, and the like. We assume that they act independently and are also on separate chromosomes and that further all plants have a homozygous

form. Thus the Bess Ross genes are represented by the following dyadic. Each x is a gene and there are N such genes.

$$x_1 x_2 \ldots\ldots x_N$$
$$x_1 x_2 \ldots\ldots x_N$$

Now we have a similar gene for the species H. multiflora. There are also N controlling independent genes and the assumption of homozygosity again holds. Thus we can write a dyadic for the species as a collection of N y genes. This species plant from which we will seek to obtain the branching is called the Non Recurrent parent. It is shown below as a dyadic.

$$y_1 y_2 \ldots\ldots y_N$$
$$y_1 y_2 \ldots\ldots y_N$$

The desired outcome is a Bess Ross but with branching. We assume that branching is dominant. If it is not then we can obtain a recessive version readily by initially backcrossing with the H multiflora and then continue as we have stipulated. The target gene structure dyadic should be as follows.

$$y_1 x_2 \ldots\ldots x_N$$
$$y_1 x_2 \ldots\ldots x_N$$

The above endpoint is what we are seeking. Backcrossing will permit this to be achieved with a high statistical probability. Namely we would obtain after a selected number of crosses the Bess Ross but with branching.

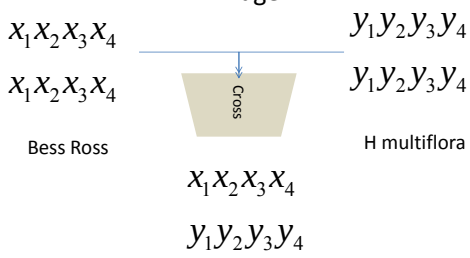
Consider a 4 Gene Case. Assume we want to insert y1 into the genome of the x sequence. Assume further that y1 is dominant. For example, we want branching from a H. multiflora to be placed into a red "Bess Ross". The Example can be generalized to N genes and even M characteristics to be "driven: in from Non Recurrent into the Recurrent.

We start by crossing Bess Ross with H multiflora. All offspring will have the genetic makeup of the following dyadic:

$$x_1 x_2 \ldots\ldots x_N$$
$$y_1 y_2 \ldots\ldots y_N$$

The F1 generation is a pure mix of the genes from both parents. We shall assume that y1 is the gene for branching and that branching is dominant. If this is not the case then we can move to F2 by crossing with H multiflora and obtain a branched sample to begin the process. We assume that there are the M genes and that each gene results in a unique expression of some phenotypic characteristic which we can measure. We could assume that there is one for color and one for branching and neglect all others. This is the more classic approach. However as we have demonstrated before, we know that there are multiple genes required and that by allowing an unspecified pool of M genes that we can achieve significantly improved results. We demonstrate this first crossing below.

The F1 cross is as follows. All F1 are identical. We assume that both initial parents are homozygous. Namely they have identical genes on both chromosomes. We further assume that there is no linkage.



In the above we assume all genes are independent and not linked. The symbolic representation is just that, a symbol for the genes not their alignment. In fact they genes may likely be on different chromosomes.

$$Let\ X = \begin{Bmatrix} x_1x_2x_3x_4 \\ x_1x_2x_3x_4 \end{Bmatrix},$$

$$Y = \begin{Bmatrix} y_1y_2y_3y_4 \\ y_1y_2y_3y_4 \end{Bmatrix},$$

$$XY = \begin{Bmatrix} x_1x_2x_3x_4 \\ x_1x_2x_3x_4 \end{Bmatrix}$$

define

$$\tilde{X}_0 = \begin{Bmatrix} x_2x_3x_4 \\ x_2x_3x_4 \end{Bmatrix}$$

$$\tilde{X}_1 = \begin{Bmatrix} x_2x_3x_4 \\ y_2x_3x_4 \end{Bmatrix} or \begin{Bmatrix} x_2x_3x_4 \\ x_2y_3x_4 \end{Bmatrix} or \begin{Bmatrix} x_2x_3x_4 \\ x_2x_3y_4 \end{Bmatrix}$$

$$\tilde{X}_2 = \begin{Bmatrix} x_2x_3x_4 \\ y_2y_3x_4 \end{Bmatrix} or \begin{Bmatrix} x_2x_3x_4 \\ y_2x_3y_4 \end{Bmatrix} or \begin{Bmatrix} x_2x_3x_4 \\ x_2y_3y_4 \end{Bmatrix}$$

$$\tilde{X}_3 = \begin{Bmatrix} x_2x_3x_4 \\ y_2y_3y_4 \end{Bmatrix}$$

Note that genes xn and yn are equally likely and have probability ½. Note that if we look at the gene tails, if they are M in length then we have $[1/2]^M$ for any one of them. Note further that for the combinations of 0, 1, 2, 3, etc we have the binomial distribution to provide the probability for any possible set of transitions from one F generation to the next F generation.

The following is a set of such transitions which are possible for this specific example. It should be readily determined what the transitions would be for any generalized form. The notation can be described as follows. If we have a cross between X0 and X0 then we can only get X0. Is we have a cross between X0 and X1, where this means that we have a tail sequence with just one y gene amongst the group, then we can get either an X0

or an X1 with equal probability. The same can then be said if we have an X0 crossed with an X2, yielding an X0, or an X1, or an X2, but now the result is controlled by a binomial distribution. The process than continues. We show the results with a three independent gene tail as follows:

$$X_0 \oplus X_0 = \{X_0$$

$$X_0 \oplus X_1 = \begin{cases} X_0; \textit{with probability 1/2} \\ X_1; \textit{with probability 1/2} \end{cases}$$

$$X_0 \oplus X_2 = \begin{cases} X_0; \textit{with proability 1/4} \\ X_1; \textit{with probability 1/2} \\ X_2; \textit{ with probability 1/4} \end{cases}$$

$$X_0 \oplus X_3 = \begin{cases} X_0; \textit{ with probability 1/8} \\ X_1; \textit{ with probability 3/8} \\ X_2; \textit{ with probability 3/8} \\ X_3; \textit{ with probability 1/8} \end{cases}$$

Now we can consider the transition from F2 to F3. Recall that F1 is merely a set of genes sharing one from each parent, the x,y combination. Then for F2, which is F1 crossed with the all X parent, we have the first form of segregation, namely we can get as the three gene tail, an all x, a one y and two x, a two y and one x, and a three y set.

To perform this analysis with a three gene tail, we will perform the analysis for each possible combination. We create a Table which shows what the crossing gene sequence is, say an X0, X1 and the like, and we then show a column which is the probability of that sequence in F2 and then we have a column for the transition probability of that sequence in F2 to the X0 sequence in F3, or the X1 sequence in F3 and so forth. This is shown below first for the X0 transition and then all others:

Cross	*Prob of This Cross in F2*	*Prob of X0 in this Cross*	*Prob X0 at F3*
X0	1/8	1	1/8
X1	3/8	½	3/16
X2	3/8	1/4	3/32
X3	1/8	1/8	1/64
Total Prob X0 in F3			27/64

Now we perform the analysis for the X1 cross elements. The second column remains the same but the third column reflects what we had demonstrated earlier. If the tail is X0 there is no chance of getting an X1 since there would be no ys available. Likewise for the X1, X2, X3 crosses we would expect a reduced number of corresponding tails in the ensuing generations.

Cross	*Prob of This Cross in F2*	*Prob of X1 in this Cross*	*Prob X1 at F3*
X0	1/8	0	0
X1	3/8	½	3/16
X2	3/8	1/2	3/16
X3	1/8	3/8	3/64
Total Prob X1 in F3			27/64

As we move to the X2 and then X3 we see that the number of them decreases at a faster rate as shown in the table below.

Cross	*Prob of This Cross in F2*	*Prob of X2 in this Cross*	*Prob X2 at F3*
X0	1/8	0	0
X1	3/8	0	0
X2	3/8	1/4	3/32
X3	1/8	3/8	3/64
Total Prob X2 in F3			9/64

Finally for X3, we see that only the tail in X3 of the prior generation do we get the chance for an X3, and that gets smaller geometrically each additional cross.

Cross	*Prob of This Cross in F2*	*Prob of X3 in this Cross*	*Prob X3 at F3*
X0	1/8	0	0
X1	3/8	0	0
X2	3/8	0	0
X3	1/8	1/8	1/64
Total Prob X3 in F3			1/64

Note that the second column is the probability of the specific sequence in F2 and that the third column is the transition probability at that specific cross to the next F generation. Namely the third column is the probability:

$$P\left[X_k(F_{n+1}) \mid X_j(F_n) \right] = p_{k,j}(n)$$

and

$$P(n) = \begin{bmatrix} p_{0,0} \cdots p_{0,N} \\ p_{N,0} \cdots p_{N,N} \end{bmatrix}$$

The above are the transition probabilities and can be readily shown to be independent of the specific crossing state, namely which Fn the probability of made for. Now we can calculate the probability of any Xn for a specific state Fk. This is as follows:

$$P[X_n(F_{k+1})] = \sum_{i=0}^{N} P[X_n(F_{k+1}) \mid X_i(F_k)]P[X_i(F_k)]$$

We have shown above that the transition probabilities are state independent and that the above equation is a recursive means to determine the next state. We demonstrate this for F4 from F3 as below:

We now do F4, and again we select the plants expressing Y1 and we again back cross with the homozygous X. This follows the same logic we did for F3. This then yields a 67% Homozygous for F4 with three genes other than the one we want impressed. The Table above can then be iterated again and again. We simply use 342/512 in the second column.

Cross	*Prob of This Cross in F3*	*Prob of X0 in this Cross*	*Prob X0 at F4*
X0	27/64	1	27/64
X1	27/64	½	27/128
X2	9/64	1/4	9/256
X3	1/64	1/8	1/512
Total Prob X0 in F4			343/512= 0.67

Cross	*Prob of This Cross in F3*	*Prob of X1 in this Cross*	*Prob X1 at F4*
X0	27/64	0	0
X1	27/64	½	27/128
X2	9/64	½	9/128
X3	1/64	3/8	3/512
Total Prob X1 in F4			147/512= 0.287

Cross	*Prob of This Cross in F3*	*Prob of X2 in this Cross*	*Prob X2 at F4*
X0	27/64	0	0
X1	27/64	0	0
X2	9/64	1/4	18/512
X3	1/64	3/8	3/512
Total Prob X2 in F4			21/512= 0.041

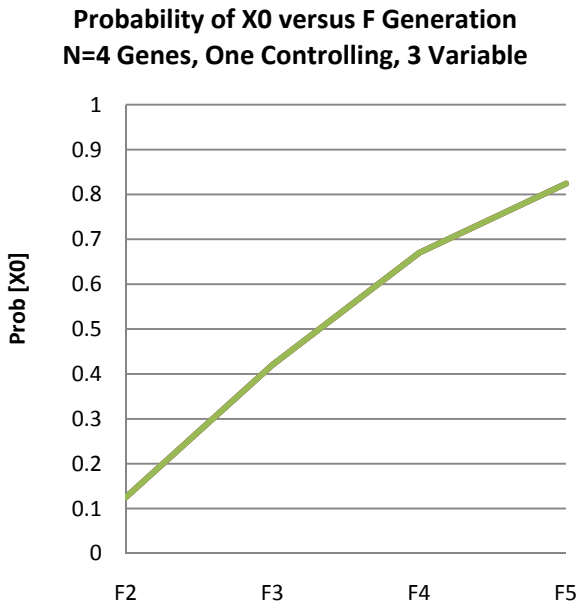
Cross	*Prob of This Cross in F3*	*Prob of X3 in this Cross*	*Prob X3 at F4*
X0	27/64	0	0
X1	23/64	0	0
X2	9/64	0	0
X3	1/64	1/8	1/512
Total Prob X3 in F4			1/512

Finally we can extend this one further time to the F5 from F4 states, focusing solely on X0. This yields the following Table using the models developed above:

Cross	*Prob of This Cross in F4*	*Prob of X0 in this Cross*	*Prob X0 at F5*
X0	0.670	1	0.670
X1	0.287	½	0.144
X2	0.041	1/4	0.010
X3	0.002	1/8	0.000
Total Prob X0 in F5			0.824

We now have a simple algorithm: The column for the last cross must be iteratively calculated for every prior step as shown. The column for the probability at the current cross can be calculated once, they will be binomial in form. The probabilities for the current and then next cross can be calculated by summing the products. Note that the larger the genome in the Recurrent the more complex and the longer the convergence.

Then we can plot the convergence rate to homozygosity in the graph shown below. Note that at F5 we have gotten to 82.4% of homozygosity.



Analyses for more complex genes and for more lengthened crossings can be accomplished. However the principle is shown in the above example. The key point to make is that the analysis we have performed herein is essential more realistic than the simplistic ones performed in the literature.

STATISTICAL ANALYSES

There are many statistical issues relating to this analysis. In this paper we focus primarily upon two issues.

First, if we assume we know M, the number of controlling genes, and we know that the model is correct, then we can determine how many crosses, N, will be required to obtain a level of selection as may be desired. One way to validate this is by testing the means of the various clusters that result and determining if they are converging at the required rate. We develop a simple test to verify this and establish bounds on the results.

Second, there is the issue of estimating the number of controlling genes, N, that may be in the backcrosses. This is a corollary to the first problem. Namely if we have two plants, each with a certain number of distinct phenotypic characteristics and we assume that we have one gene and one phenotype, then the question is how many genes are in this backcross mix? We have assumed that we know N, the number of genes. In reality we most likely do not know N, however we know the number of generations by definition, we have measures on the phenotype characteristics and their respective frequency. Thus we should have enough to obtain an estimate of N by using the assume convergence model developed herein, and furthermore we can obtain bounds on the accuracy of the estimate of the value of N obtained thereby.

We first consider the question of how many generations we must cross to attain a desired level of homozygosity. We know from classic t-statistics how to size and experiment for a specified level of certainty if we were to see if the mean were within certain bounds and within the desired level of certainty. There are also simple tests to determined paired samples. However in this case the problem can be stated more complexly. We have N characteristics and we know what the means are for the number of samples in each of the characteristic sets. We further know that as we increase the number of crosses M to a larger number that the average number in the sets being crossed against decrease exponentially. In reality we only desire to retain the set for which we are backcrossing and whose presence is exponentially increasing. Thus the determination of the number of samples required to reach a level of confidence may be obtained by focusing on the X0 set only and then doing so in each Fn generation (see Pagano and Gauvreau).

We can now address the second issue. Namely, given a set of sequential measurements of phenotypes, what is a reasonable estimator of M, the number of genes controlling the phenotypes. Consider the following experiment. Let n be the nth cross, with corresponding generation Fn. Let there be a total of N such generations. Let B be any resulting set of normalized results for a phenotype in that generation. We will detail this as follows:

$$B_k(n) = \frac{T_k(n)}{\sum_{i=1}^{M} T_i(n)}$$

where $T_k(n)$ is the total with phenotype i at Fn

Now we know that:

$$P[X_k(n+1)] = \sum_{j=0}^{M} p_{k,j} P[X_j(n)]$$

Which we can write as:

$$T_k(n) = T(n)P[X_k(n)]$$

and T(n) is the total number in the Fn generation

Now we can also look at each of the values of T or equivalently the normalized values we define as B, as follows.

$$P[M \mid B] = \frac{P[B \mid M]P[M]}{P[B]}$$

where;

$$B = \{B_0(1)....B_M(N)\}$$

But we also can say that:

$$B_k(n) = \bar{B}'_k(n) + w_k(n)$$

where

$$\bar{B}'_k(n) = P[X_k(n)]$$

We can use a maximum likelihood estimator which gives M as follows:

Find M to maximize:

$P[B|M] =$

$P[B_0(1)...B_M(1)...B_0(N)....B_M N)\,|\,M]$

Now we can use the previous observation to state that the Bs have known means, given M, and that we can calculate them, and that they are random variables with w being a zero mean Gaussian with variance σ and we can further assume that they are independent. Then using the log of the likelihood function as defined we can then obtain an estimator which minimizes that sum of squares. Now we need to determine the variances on each of the samples. The variances will be used to weight each sample. Before proceeding we can restate the ML solution as follows:

Find M to minimize:

$$\sum_{n=1}^{N}\sum_{m=0}^{M}\frac{(\bar{B}_m(n)-B_m(n))^2}{\sigma_m^2(n)}$$

We can use the sample variances for the ensemble variances. Similarly we can calculate the ensemble variances using the fact that the ensembles are generated by the binomial selection processes. The ensemble variances are quite difficult to calculate so we retain the sample variances as simpler measures.

Now we can determine the variance on the estimate by using the Cramer-Rao bound which functions well on such Gaussian analyses (see Van Trees). Specifically we have:

$$\mathrm{var}(M-\hat{M}) \geq \left(E\left\{\left[\frac{\partial^2 \ln p(B\,|\,M)}{\partial M^2}\right]\right\}^{-1}\right)$$

But since these variables are assumed Gaussian this can be calculated readily for any M.

As an example, we could consider the crosses we had discussed above. If we look at Bess Ross and H multiflora, we could consider 2 genes, color and branching, and then go from there. For three genes, we could introduce the root, tubular versus bulbous, then length of scape, length of leaf, width of leaf, number of flowers per branch, and so forth. We note that as we increase the number of putative genes, the denominator which represents the total number of samples, goes up, driving the ratios for each gene down. As we increase the genes we then get more variation and it goes up again. Thus, arguably there is a minimum.

The method proposed is actually a form of cluster analysis (see Fukunaga). It seeks to find the optimal number of clusters of values for sets of characteristics. By examining the method, the clusters are based upon a collection of characters. For example, if we have N=2, then for all branched plants we have color and scape length as possible characters. We then sort on the four possible sets; red and long scape, red and short scape, yellow and long scape and yellow and short scape. The Bess Ross could be defined as red and short scaped. We could then also expand it to the other characteristics as we have discussed before.

DISCUSSION

The ability to backcross is an essential element in hybridizing. It permits the introduction of a trait into an existing line and then ensuring that the line is returned to its original genetic state with the exception of the new phenotypic characteristic having been expressed. All other phenotypic characteristics are returned to where they were at the initial state.

There are several additional enhancements which must be made to this analysis. First, linkages must be incorporated. For F1 through typically F5 the linkages of genes may not play a significant role. However as we continue to backcross there are increasingly import effects of linkages which must be accounted for (see Griffiths). Second, we know that many of the genes are modulated by repressor and activator genes. These must also somehow be accounted for. Generally, if they are not affecting other genes we can let them be second order effects. However, when they cross modulate in gene expression motifs then we have to establish their presence in the model. Third, this is an analysis and hybridizing planning tool. This is not a synthesis tool as currently structured.

LITERATURE CITED

Allard, R., **Plant Breeding**, Wiley (New York) 1960.

Brown, A., et al, **Plant Population Genetics, Breeding, and Genetic Resources**, Sinauer (Sunderland, MA) 1990.

Campbell, A., L. Heyer, **Genomics, Proteomics, and Bioinformatics**, Benjamin Cummings (New York) 2003.

Crawford, D., **Plant Molecular Systematics**, Wiley (New York) 1990.

Cronquist, A., **The Evolution and Classification of Flowering Plants**, New York Botanical Garden Pres (Bronx, NY) 1986.

Dunn, G., B. Everitt, **Mathematical Taxonomy**, Dover (Mineola, NY) 2004.

Durbin M. L. et al, Genes *That Determine Flower Color*, *Molecular Phylogenetics and Evolution*, 2003 pp. 507-518.

Erhardt, W., **Hemerocalis**, Timber Press (Portland, OR) 1992.

Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic (Boston) 1990.

Griffiths, A., et al, **Genetic Analysis 5th Ed**, Freeman (New York) 1993.

Gusfield, D., **Algorithms on Strings, Trees, and Sequences**, Cambridge (New York) 1997.

Hildebrand, F. B., **Numerical Analysis**, 2nd Edition, Dover (New York) 1987.

Judd, W., et al, **Plant Systematics**, 3rd Ed, Sinauer (Sunderland, MA) 2008.

Mayo, O., **The Theory of Plant Breeding**, Oxford (New York) 1987.

McGarty, T. P., *Flower Color and Means to Determine Causal Anthocyanins And Their Concentrations*, MIT 2008, http://www.telmarcgardens.com/Documents%20Papers/Flower%20Color%20and%20Means%20to%20Determine%2002.pdf

McGarty, T. P., *On the Structure of Random Fields Generated by a Multiple Scatter Medium*, PhD Thesis, MIT 1971. http://mit.edu/mcgarty/www/MIT/Paper%20Hypertext/1971%20PhD%20MIT.pdf

McGarty, T., *Gene Expression in Plants*: Use of System Identification for Control of Color, MIT, 2007.

http://mit.edu/mcgarty/www/MIT/Paper%20Hypertext/2007%20Gene%20Expression%20IEEE%2007%2002.pdf .
McGarty, T., **Stochastic Systems and State Estimation**, Wiley (New York) 1974.
Murray, J., **Mathematical Biology**, Springer (New York) 1989.
Murrell, J., *Understanding Rate of Chemical Reactions*, University of Sussex.
Nei, M., S.. Kumar, **Molecular Evolution and Phylogenetics**, Oxford (New York) 2000.
Norton, J., *Some Basic Hemerocallis Genetics*, American Hemerocallis Society, 1982.
Pagano, M., K. Gauvreau, Principles of Biostatistics, Duxbury (Belmont, CA), 1993.
Percus, J., **Mathematics of Genome Analysis**, Cambridge (New York) 2002.
Sokal, R., P. Sneath, **Principle of Numerical Taxonomy**, Freeman (San Francisco) 1963.
Stout, A.B., **Daylilies**, Saga Press (Millwood, NY) 1986.
Strickberger, M., **Genetics**, 2nd Ed McMillan (New York) 1976.
Studier, J., K. Kappler, *A Note on the Neighbor Joining Algorithm*, Molecular Biological Evolution, Vol 5 1988 pp 729-731.
Stuessy, T., **Case Studies in Plant Taxonomy**, Columbia University Press (New York) 1994.
Taiz, L., E. Zeiger, **Plant Physiology**, Benjamin Cummings (Redwood City, CA) 1991.
Taubes, C. H., Modeling Differential Equations in Biology, Cambridge (New York) 2001.
Van Trees, H. L., **Detection, Estimation and Modulation Theory**, Wiley (New York) 1968.
Watson, J., et al, **Molecular Biology of the Gene**, Benjamin Cummings (San Francisco) 2004.
Weir, B., **Genetic Data Analysis**, Sinauer (Sunderland, MA) 1990.