Phylogenetics, DNA, Classification And the Genus Hemerocallis[1]

Terrence P McGarty

Abstract

The genus Hemerocallis has had various attempts at classification since the time of Stout. The primary approach has been via phenotypic methods and Erhart has recently proposed an alternative classification to Stout. With the introduction of various genetic methods for classifying the genus, a dendrogram has been proposed using the AFLP methods of genetic sampling. This paper reviews the various techniques and then also reviews and summarizes the several phylogenetic approaches which have been taken to the present. The paper then details some analyses of comparisons of the papers in the literature and makes suggestions as to potential next steps. To data there has not been a detailed genetic analysis which has allowed for detailed evolutionary classification.

Content

[1] Copyright © Terrence P McGarty 2008 all rights reserved. tmcgarty@telmarc,com , www.telmarc.com Charts and Tables have been used and their source has been referenced. This is covered under the fair use doctrine. The material contained herein is the opinion of the author and no representation is made as to its accuracy. Reliance upon material contained herein is at the sole risk of the user.

1 INTRODUCTION

The classification of plants involves many complex methodologies. Recently, with the use of DNA methodologies, there has been a re-assessment of many of the classifications based upon morphology and the other more classic limited metrics and measurements. This paper focuses on classifications using those newer techniques. It looks at what the methodologies are and evaluates their respective advantages and disadvantages.

It then looks at the methodologies of taking the data collected from the DNA methodologies and then creating classifications based upon that data. Certain methods use generally gross level methods and others use quite detailed and sophisticated methodologies which attempt to incorporate actual DNA modification. The eventual goal of this paper is the set down a classification of the genus Hemerocallis which reflects the best current thinking using DNA measurements.

We begin the discussion with the posing of several key questions and then we try to answer them with the tools currently available. There are many such questions which can be easily posed but not readily answered. The genus Hemerocallis is a simple genus of a monocot plant which is originally from Asia, including China, Japan and Korea. It is a common plant which comes in a variety of species.

1.1 Key Questions

There are many questions which beg the answering. The following are several:

1. What are the species in the genus Hemerocallis? In fact, what do we mean by species?

Ernst Mayr was famous for defining a species as a collection of living organisms which have the capability of interbreeding. (See the various works of Mayr on the issue of species) Elephants and lions do not interbreed, thus it is obvious that they are different species. The genus Pinus and the genus Picea do not interbreed, thus they are composed of different species. Yet the species of Hemerocallis readily interbreed, begging the question of species as posed by Mayr. Hemerocallis all have 11 pairs of chromosomes, namely they are diploid with a total of 22 chromosomes, with the exception H fulva whish is a diploid with a total of 33 chromosomes.

2. Within what one may see as a species, how much variation can one tolerate and still call it a species?

This is a key question. Is there a specific characteristic which defines a species? For example, H dumortierii has brownish sepals, and it is sessile. If the sepals are no longer brown is this now a new species? Or is it just a variation? What is controlling the color, is it a definite species characteristic?

3. What characteristics do we look for to distinguish a species one from the other? What are the most telling of the characteristics, and why does one select those characteristics?

When we look at H fulva we know it has 33 chromosomes. That makes H fulva unique. It also is generally sterile. Then we have species which all bloom at about the same time. The H minor, H dumortieri H middendorfii and H flava all bloom at the same time. Perhaps this means that they could inbreed. But dumortierii and middendorfii have sessile flowers whereas minor and flava have branched flowers. Is sessile and branching a major factor which makes them species? H minor is grass like, with drooping scapes and leaves which droop and are thin and short. H flava is erect with larger leaves. Both naturally pollinate by bees. However flava blooms at night and minor is more of a day bloomer.

4. If we can identify a species, and we can see the collection of all species, how can we relate one species to another? Is there some closeness of one species with another, and moreover is there a way to relate them so as to see how they evolved to where they are now? Finally, can we "look back in time" to understand what the ancestor was or the ancestors were?

This is the process of developing a tree showing the relationship of one species to another. The issues of defining the relationship are driven by a closeness measure. It can also be driven by a change in genes. For example if one species has a gene given by:

…CCTTAGCCT…

And the other species has a gene:

…CCATAGCCA…

Then we may ask what ways did these genes get to this point? If we know that genes mutate at the rate of α per thousand years then we could calculate the most likely ancestor of these two genes. This is one of the many ways one can approach this problem. At the heart of any such approach is some measure of closeness. How close are the two genes, how close are two proteins, and so on.

5. Using genetic tools, how would we best approach the issues of identifying species? What are the best genetic markers, and how detailed should one get to optimize the task? Given the best possible genetic marker, how do we then sort and arrange the measurements to assist in defining species?

There are thousands of genes. Which ones should we focus on and should we weight them differently and if so how differently? One can assume that they have the best set of genes from all the species. Then one must look at both intra species matches and interspecies matches.

This paper examines many of these questions. There are answers for some, work in progress for others, and many which are still a way from being addressed.

1.2 Prior Efforts

In the past ten years there have been many studies regarding the genus Hemerocallis. We briefly review a few.

Chung and Noguchi in 1998 published a paper on H middendorffii where they looked at the differences in morphological characteristics over regions in Japan and Korea. This paper provides a good benchmark for the use of morphology. It shows that there is also some significant variation within a species as to the morphological characteristics.

Chung in 2000 collected H hakuunensis samples and using an enzyme technique examined the spatial variability within the species. Three specific enzymes were analyzed and there was significant spatial variation was found in one and little in two others. The species has some variability but not a great deal.

Hasegawa et al in 2006 reported on hybridization between H fulva and H citrina. The fulva is a day blooming plant and citrina a night bloomer. There is some crossing that result from the slight overlap of bloom. Specifically the authors' state: "*Most F1 hybrids showed diurnal flowering. These findings indicate that only a few genes have strong phenotypic effect on the determination of lowering time in Hemerocallis, and suggest that the evolution from a H. fulva-like ancestor to H. citrina was not a continuous process by accumulation of minute mutations*." This study has been flowed up by Yasumoto and Yahara in 2008 where they deliberately set F1 crosses. The belief is that H fulva is an ancestor to H citrina.

The work by Kang and Chung in 1997 examined the genetic variation in H hakuunensis. The authors used enzyme markers and they observed:

"Hemerocallis hakuunensis, a Korean endemic species, maintains considerably higher levels of allozyme variation within populations …and substantially lower levels of allozyme divergence among populations ….. than average values reported for other insect-pollinated, outcrossing herbs. Indirect estimates of the number of migrants per generation … indicate that gene flow has been extensive in H. hakuunensis. This is somewhat surprising when we consider the fact that no specialized seed dispersal mechanism is known, flowers are visited by bees, and the present-day populations of the species are discontinuous and isolated. Results of a spatial autocorrelation analysis based on mean allele frequencies of 19 populations reveal that only 13% … of Moran's I values for the ten interpopulational distance classes are significantly different from the expected values and no distinct trend with respect to the distance classes is detected. Although it is unclear how the populations are genetically homogenous, it is highly probable that H. hakuunensis might have a history of relatively large, continuous populations that had more chance for gene movement among adjacent populations after the last Ice Age. In addition, occasional hybridization with H. thunbergii in areas of sympatry in the central and southwestern Korean Peninsula may be one factor contributing the present-day high allozyme variation observed in H. hakuunensis."

The Kang and Chung study is one of the first to detail genetic markers.

Kang and Chung in 2000 looked at the high levels of enzyme variation within a population and low divergence within and amongst species. This was done for H. thunbergii, hakuunensis, hongdoensis, taeanensis, middendorffii, thunbergii. Specifically the authors' state:

"Hemerocallis thunbergii, H. hakuunensis, H. middendorffii, and H. taeanensis had high genetic diversity. On the other hand, three populations of H. hongdoensis maintained significantly … lower mean values of HEe…. than those for the other four Hemerocallis species. Hemerocallis hongdoensis also had the lowest number of alleles…"

"As expected, Korean populations of H. thunbergii and H. middendorffii have high genetic diversity. The two species have a wide geographic range distributed from China to parts of the Korean Peninsula and the Japanese Archipelagos. Most Korean populations of H. thunbergii grow commonly in the open, grasslands on hillsides in the southwestern Korean Peninsula. It has been observed that Korean populations of the species are large and have a relatively continuous distribution."

From the Kang and Chung paper they provide a classification based upon the enzyme studies as follows:

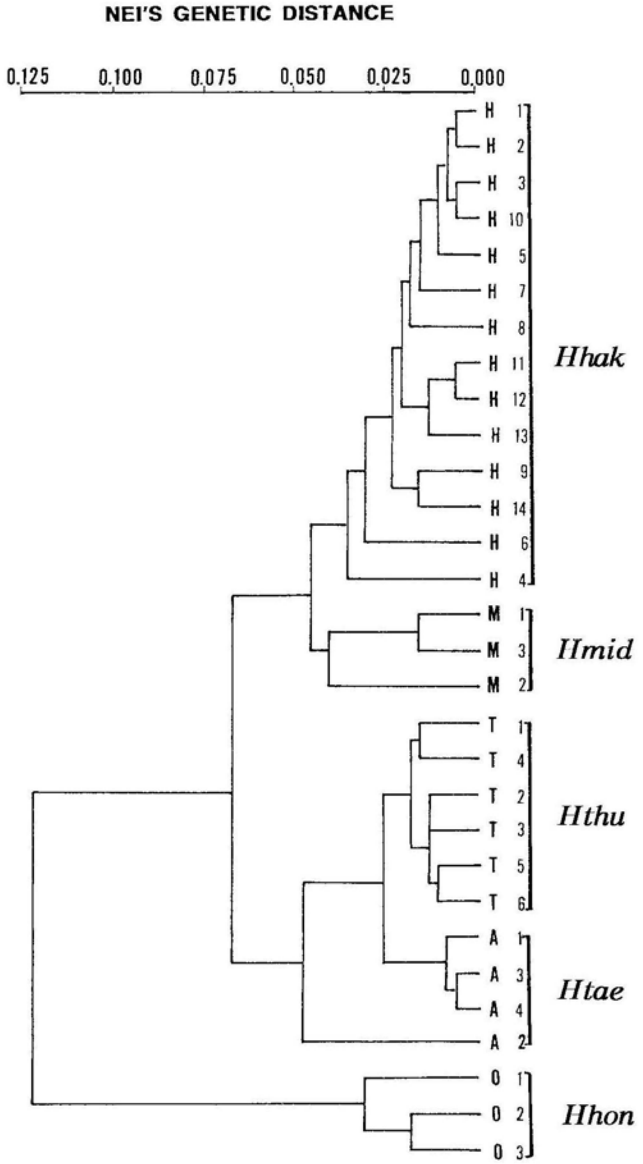


Fig. 3. UPGMA cluster analysis of 30 populations of *Hemerocallis* species in Korea based on Nei's (1972) measure of genetic distance. Abbreviations are from Fig. 1 and Table 3. *Hhak, H. hakuunensis*; *Hmid, H. middendorffii*; *Hthu, H. thunbergii*; *Htae, H. taeanensis*; and *Hhon, H. hongdoensis.*

In the above classification, what are grouped are the intraspecies and the interspecies. The grouping methodology, UPGMA, is discussed herein. The classification demonstrates several key facts:

1. Intraspecies variation can be significant. In the above we have five species and we can see that the H hakunensis has substantial intra species variability.

2. Interspecies variation is also quite extensive. It is not at all clear from this dendrogram how far back in evolutionary time the species split but we can see that H middendorfii and H hakunensis

are related as are H thunbergii and H taeanensis, whereas H hongdoensis is not. The question then is which is closest to the true ancestor.

Tompkins et al in 2001 published the first paper on the use of AFLPs to determine the genetic variation in Hemerocallis. We will focus on their work latter in the paper. Their study presents one of the first truly comprehensive genetic dendrograms or classifications of the genus.

Guerro et al in 1998 performed a detailed genetic analysis of the specific genes which controlled senescence. They used cDNA genes for this specific purpose. This appears to be one of the first truly gene studies and one of the first to create cDNA for the genus.

2 THE PROBLEM OF CLASSIFICATION

Classification of species has been at the heart of all plant systematics. The classification process generally tries to arrange plants into a logical form and doing so to sort the species in some evolutionary manner. Thus the magnolia is a more distant entry into the angiosperms and the asters are more recent. This conclusion is based upon the appearance of certain morphological characteristics found in what may now be extinct plants. One may see a certain characteristics in a magnolia which is found at period X and then see the aster characteristic in period Y and Y is more recent than X and thus the asters are in an evolutionary sense a more recent group then the magnolias. This is a simplistic way to explain the process.

This type of classification works well on families and possibly on genera, if at all. It seems not to work well on species because the historical evolutionary evidence is lacking. Thus species are related purely by the current morphological characteristics.

In the genus Hemerocallis one common characteristic could be sessile flowers versus branched flowers, an approach taken by Stout. At the other extreme would be the analysis of the genes of various species and then to attempt to relate one to the other.

The issue of genetic relating can be phrased as follows:

1. The genus Hemerocallis has 11 chromosomes and 22 chromosome pairs in the diploid species. There is estimated to be several thousands of genes, and the gene length may vary from dozens to hundreds of nucleotides.

2. Certain of the genes have been identified and certain of them are common across other families, such as the genes controlling the secondary pathways of the pigment sources.

3. If we were to look at a large enough collection of genes, and then compare them both within and between species it would be possible to characterize the species based upon the genetic consistency. Current methods in bioinformatics would allow for the assessment of consistency across the gene structures.

4. Using tools that have been developed which account for the changing of genes due to various mechanism one may be able to take the set of existing species and then work backward to

attempt to determine how the speciation occurred genetically and how long such speciation may have take and also to determine if there was one or several ancestors. This can be accomplished using the maximum likelihood approach which we discuss herein. Such an approach is highly complex.

Thus it is possible that in time the genetic speciation of Hemerocallis can be elucidated. The state of knowledge at the current time however does not permit that. There has been a great deal of work using other methods which we discuss herein.

One of the important issues to address when performing a phylogenetic assessment is to clearly delineate between intra-species and inter-species variations. In the dendrogram shown below from the work of Kang and Chung (1997) the authors genetically analyzed the species H. hakuunensis and from that analysis demonstrated significant genetic variation within the species. The authors' state:

"Hemerocallis hakuunensis, a Korean endemic species, maintains considerably higher levels of allozyme variation within populations… and substantially lower levels of allozyme divergence among populations… than average values reported for other insect-pollinated, outcrossing herbs. Indirect estimates of the number of migrants per generation … indicate that gene flow has been extensive in H. hakuunensis. This is somewhat surprising when we consider the fact that no specialized seed dispersal mechanism is known, flowers are visited by bees, and the present-day populations of the species are discontinuous and isolated."

The authors used enzyme analysis to develop the following tree.

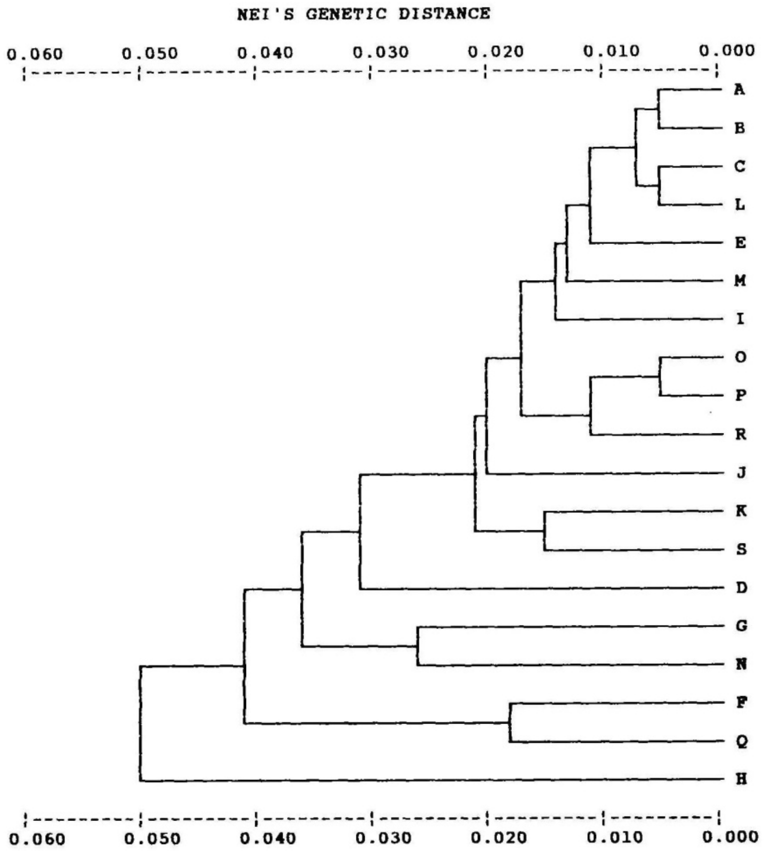


Fig. 2. Dendrogram from UPGMA cluster analysis based on Nei's (1972) genetic distance between the 19 populations of *Hemerocallis hakuunensis*.

The tree show demonstrates a significant intra species variation of the limited enzymes being assayed. It generally is consistent with what we have shown before. Thus one is led to assume that if one looked at the genetic variation across a large base of genes that the variation within species would be significant.

2.1 Morphological Classification

The classic approach to classification has been to use plant morphology. The use of such factors as those in the Table below has been done by many authors including those we have summarized in the introduction. Whether these are the best set are open to discussion.

Traits
No. of scapes (#)
Flower tube length (mm)
Petal length (mm)
Petal width (mm)
Stamen length (mm)
Pistil length (mm)
Plant (scape) height
Length of inflorescence minus flowers (cm)
Length of the lowest bracts (cm)
Number of flowers/scape (#)
Length of the perianth tube enclosing an ovary (cm)
Length of the outer perianth (cm)
Width of the outer perianth (cm)
Length of the inner perianth (cm)
Width of the inner perianth (cm)
Length of the widest leaf (cm)
Width of the widest leaf (cm)

We shall consider in detail the use of morphology in another paper. However it is worth considering what two people have done in the past one hundred years. In 1934 Stout published his book on daylilies. This was the first work and it was a work prepared by one skilled in the art. He was both a PhD in the field and he had even by then been active at the New York Botanical Garden, then and now a pre-eminent institution in the botanical area. He associated with Cronquist and others who had a great impact on the development of systematics. In his book he proposed a key to the species. It was a key, NOT a phylogeny. It was to be used to identify the species and NOT to specify any evolutionary or genetic relationship. He simply broke the species into two classes, those with branches and those without. Then he went down from there. Given what he had to work with, albeit extensive, he had not yet been able to identify all species and he did not have the advantage of thousands of others in the field.

In 1992 Erhardt in his book on Hemerocallis proposed a Classification, not a Key. The term classification carries a great deal more weight than a key. Keys help identify and classifications establish relationships. Erhardt states in his book:

"Stout's proposed division was not accepted and no one now supports it>"

Frankly that statement is a combination of arrogance and ignorance. By its face it uses the term division, not Key and not Classification. Division as a term of art has no standing. In addition if it was unused and in fact as implied by Erhardt was useless then why no one did from 1934 until Erhardt in 1992, sixty years, ever propose another, if we are to believe Erhardt. In fact there were dozens of others, all with slight nuances as new data was determined. Erhardt goes on:

"In my view there are five main groups of the day lily and the members in each group are either related or are perhaps varieties of one another."

Erhardt is a self declared "plantsman with wide ranging horticultural interests…" He clearly seems to lack the academic training given his self representation and one must ask what the basis for his selection was. There is no justification, just a statement of what he perceives as a fact. However, it is worth the exercise to examine his five morphological groups.

1. Fulva Group: Blooms are reddish, roots are bulb like, and this contains H fulva and H aurantiaca.

2. Citrina Group: Blooms mostly yellow, long perianth tubes and bloom opens in evening. The scapes are branched: H altissima, H citrina, H coreana, H lilioasphodelus (flava), H minor, H thunbergii.

3. Middendorffii Group: Blooms are orange, and they are sessile. Bracts are short and overlap: H dumortieri, H hakunensis, and H middendorfii.

4. Nana Group: Short scapes short perianth, not winter hardy: H forrestii, H nana.

5. Multiflora Group: Flowers on short stalks, branched, smaller flowers: H multiflora.

He then uses this classification to generate a Key. Thus he clearly knows or should know what the difference between a Classification and a Key is. Furthermore when he characterizes H hakunensis he says it is branched. Well it is or it is not. This is typical or Erhardt.

A better approach would be to look at the characteristics and see how they evolved. As we indicated earlier there is some evolutionary evidence to attest to the fact that citrina came from fulva. Then are all in the citrina group related, because of the night blooming. H minor is a very early bloomer, whereas citrina is later, typically four to six weeks, and coreana is even later. H altissima is very late and is a tall plant. They are all fragrant and one can hybridize between with some success.

In contrast H dumortieri and H middendorffii are both sessile, blooms at the same time, and seem to be cross sterile. Are these genetic variants of one another? On the other hand H hakunensis is a late bloomer and is not really sessile. Thus what was Erhardt's basis for making these aggregations?

We will examine this latter in the paper.

2.2 Genetic Classification

Recently, during the past twenty years, there has been a massive development in amount of tools available to both collect and analyze genetic data. Collection methods of proteins, enzymes, DNA, mRNA, cDNA, and variants of these have been developed. From the simple and now classic Southern blots to the use of million cell microarrays we now have a vast collection of raw genetic data potentially available.

The processing of this data for the genus Hemerocallis has just commenced. In addition to the collection techniques there are in many ways even more in terms of analytical tools. The tools range from the complex mechanism which align genes and search for genetic patterns, to those which take those patterns and use sophisticated statistical models to reverse engineer the evolutionary changes.

Techniques like the maximum likelihood technique have been used in communications for decoding signals which have been coded and sent across noisy and dispersive channels. The same processing used in this communications applications are now used in genetic analysis

3 GENETIC TECHNIQUES

In this section we provide a summary of the key genetic techniques we need to understand in order to approach systematic from a genetic perspective.

3.1 Genes and Restriction Enzymes

This first section reviews several of the essential elements we need to take the next step and select genes and proteins. There are two elements we review; restriction enzymes and polymerase chain reactions.

3.1.1 Restriction Enzymes

One of the earliest discoveries in understanding DNA and genes was the recognition that certain enzymes, proteins, have the ability to cut DNA at certain well defined points in a consistent manner. These enzymes are called restriction enzymes and they allow one to select areas for cutting.

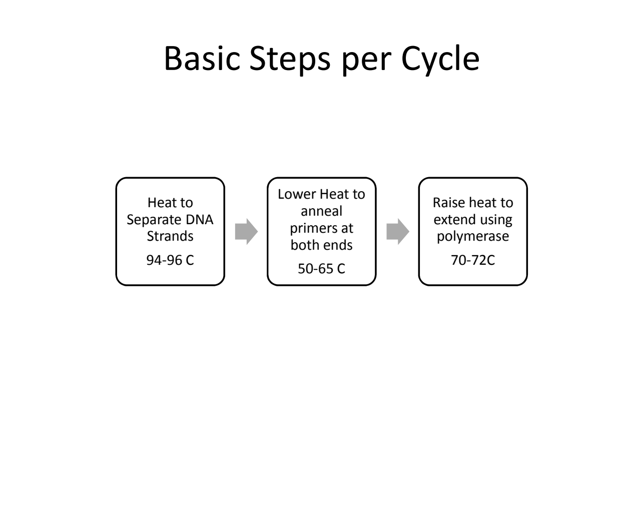
The following table is a list of some of the most important restriction enzymes. The Table lists the name of the enzyme, its source, namely what organism it has been obtained from, the target sequence it finds to cut at and the cut sequencing.

Restriction enzymes allow one to take long strands of DNA and to cut them in a predictable manner. Having these predictable cuts we can now add tags to the strands or do whatever else we seek to do.

Enzyme	Source	Recognition Sequence	Cut
EcoRI	Escherichia coli	5'GAATTC3'CTTAAG	5'---G AATTC---3'3'---CTTAA G---5'
EcoRII	Escherichia coli	5'CCWGG3'GGWCC	5'--- CCWGG---3'3'---GGWCC ---5'
BamHI	Bacillus amyloliquefaciens	5'GGATCC3'CCTAGG	5'---G GATCC---3'3'---CCTAG G---5'
HindIII	Haemophilus influenzae	5'AAGCTT3'TTCGAA	5'---A AGCTT---3'3'---TTCGA A---5'
TaqI	Thermus aquaticus	5'TCGA3'AGCT	5'---T CGA---3'3'---AGC T---5'
NotI	Nocardia otitidis	5'GCGGCCGC3'CGCCGGCG	5'---GC GGCCGC---3'3'---CGCCGG CG---5'
HinfI	Haemophilus influenzae	5'GANTC3'CTNAG	5'---G ANTC---3'3'---CTNA G---5'
Sau3A	Staphylococcus aureus	5'GATC3'CTAG	5'--- GATC---3'3'---CTAG ---3'
PovII*	Proteus vulgaris	5'CAGCTG3'GTCGAC	5'---CAG CTG---3'3'---GTC GAC---5'
SmaI*	Serratia marcescens	5'CCCGGG3'GGGCCC	5'---CCC GGG---3'3'---GGG CCC---5'
HaeIII*	Haemophilus aegyptius	5'GGCC3'CCGG	5'---GG CC---3'3'---CC GG---5'
AluI*	Arthrobacter luteus	5'AGCT3'TCGA	5'---AG CT---3'3'---TC GA---5'
EcoRV*	Escherichia coli	5'GATATC3'CTATAG	5'---GAT ATC---3'3'---CTA TAG---5'
KpnI[1]	Klebsiella pneumoniae	5'GGTACC3'CCATGG	5'---GGTAC C---3'3'---C CATGG---5'
PstI[1]	Providencia stuartii	5'CTGCAG3'GACGTC	5'---CTGCA G---3'3'---G ACGTC---5'
SacI[1]	Streptomyces achromogenes	5'GAGCTC3'CTCGAG	5'---GAGCT C---3'3'---C TCGAG---5'
SalI[1]	Streptomyces albus	5'GTCGAC3'CAGCTG	5'---G TCGAC---3'3'---CAGCT G---5'
ScaI[1]	Streptomyces caespitosus	5'AGTACT3'TCATGA	5'---AGT ACT---3'3'---TCA TGA---5'
SphI[1]	Streptomyces phaeochromogenes	5'GCATGC3'CGTACG	5'---G CATGC---3'3'---CGTAC G---5'
XbaI[1]	Xanthomonas badrii	5'TCTAGA3'AGATCT	5'---T CTAGA---3'3'---AGATC T---5'

3.1.2 PCR

The polymerase chain reaction, PCR, was a brilliant step in the management of DNA. It allowed for the multiplication of small snippets of DNA into millions of copies of the small snippet. The process is shown at high level below. It goes through three heat stages, heat to break DNA apart, then cool to bond a marker, then heat again to get the enzymes to build out the DNA again along the new track created by the market. The separate, anneal and extend process is copied over and over.

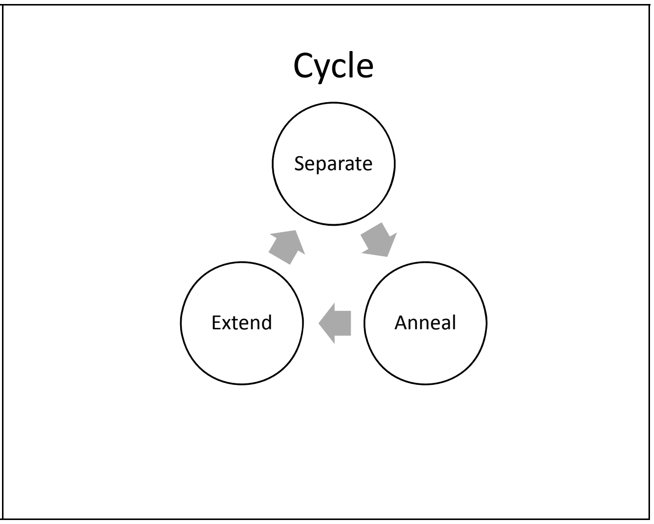


Basic Steps per Cycle

Heat to Separate DNA Strands 94-96 C

Lower Heat to anneal primers at both ends 50-65 C

Raise heat to extend using polymerase 70-72C



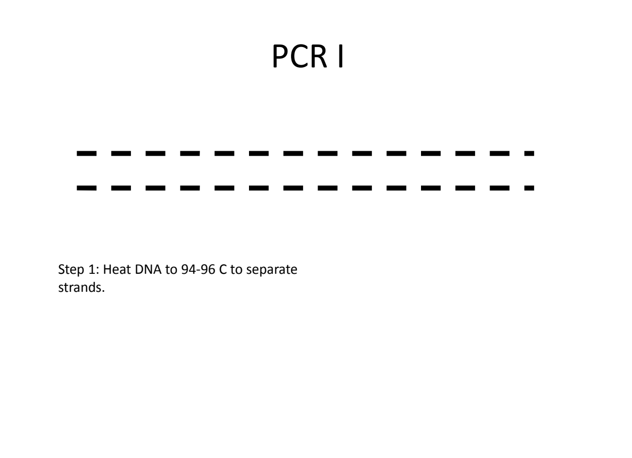
Cycle

Separate

Anneal

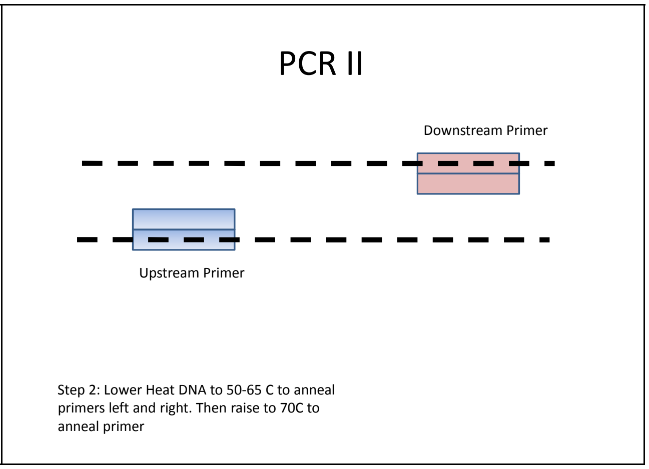
Extend

The specific details are shown in the following ten steps. Simply stated we separate, anneal a marker, re-grow the DNA now with the marker, repeat this with new markers, so that by the third step we have the segments with two end markers making themselves over and over, and they have exponential growth. Ten cycles, we get 2 to the 10th and this is a thousand fold multiplications, twenty cycles we have millions, all from a single strand!



PCR I

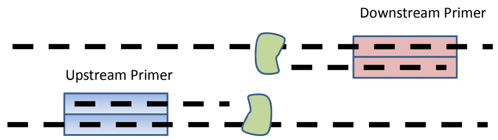
Step 1: Heat DNA to 94-96 C to separate strands.



PCR II

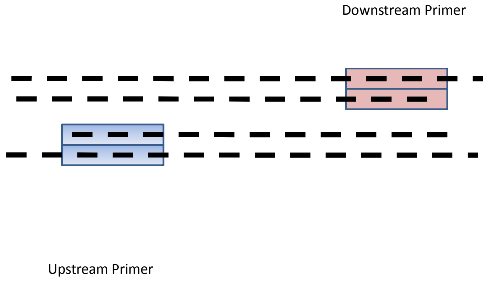
Step 2: Lower Heat DNA to 50-65 C to anneal primers left and right. Then raise to 70C to anneal primer

PCR III



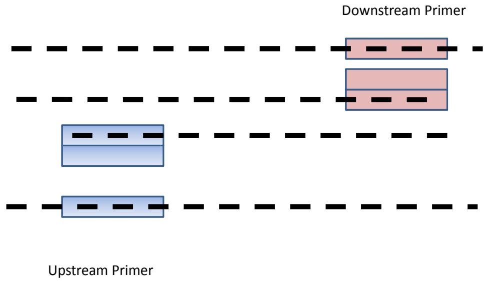
Step 3 At 72 C use polymerase to extend Primers

PCR IV



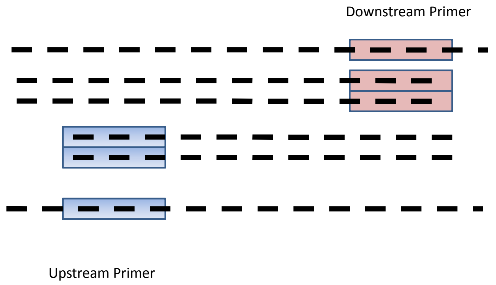
Step 4 Start Cycle 2

PCR V



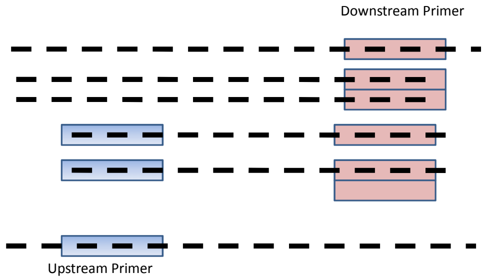
Step 4 Start Cycle 2

PCR VI



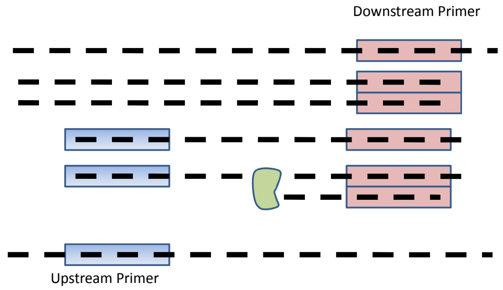
Step 4 Start Cycle 2

PCR VII

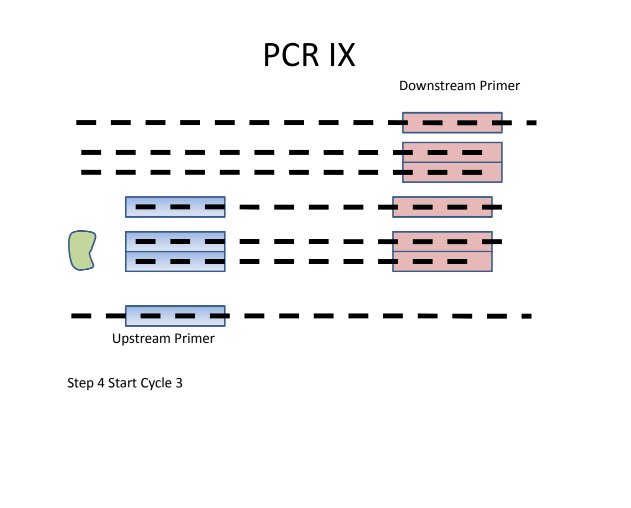


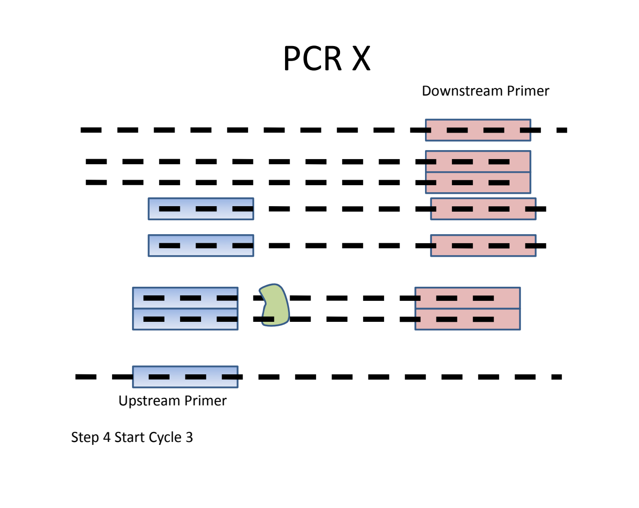
Step 4 Start Cycle 2

PCR VIII

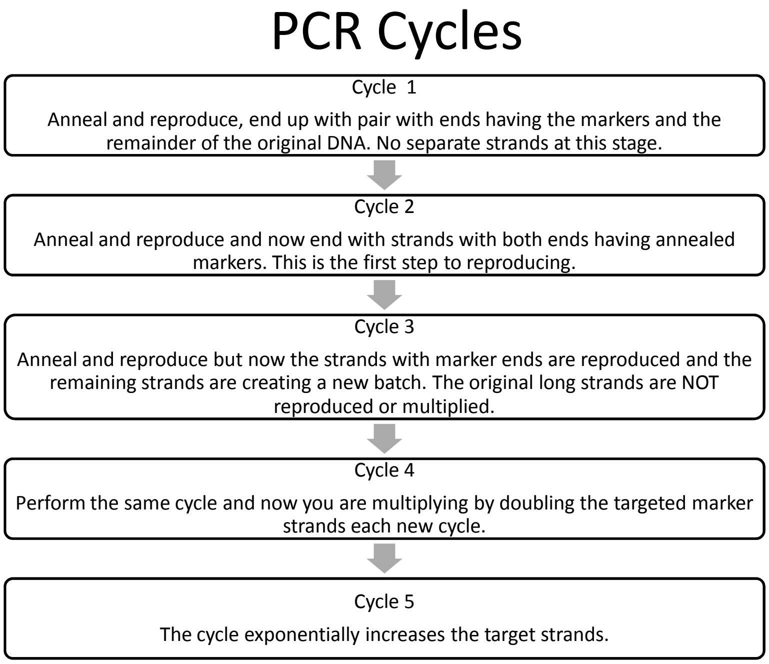


Step 4 Start Cycle 2





The following Table summarizes the cycles which we have shown above in extreme detail. The cycles require the first three to obtain a double ended segment and from that point on that specific segment is doubled at each part of the PCR cycle. There are also PCR systems which perform this cycling on a continuing basis.



PCR Cycles

Cycle 1

Anneal and reproduce, end up with pair with ends having the markers and the remainder of the original DNA. No separate strands at this stage.

Cycle 2

Anneal and reproduce and now end with strands with both ends having annealed markers. This is the first step to reproducing.

Cycle 3

Anneal and reproduce but now the strands with marker ends are reproduced and the remaining strands are creating a new batch. The original long strands are NOT reproduced or multiplied.

Cycle 4

Perform the same cycle and now you are multiplying by doubling the targeted marker strands each new cycle.

Cycle 5

The cycle exponentially increases the target strands.

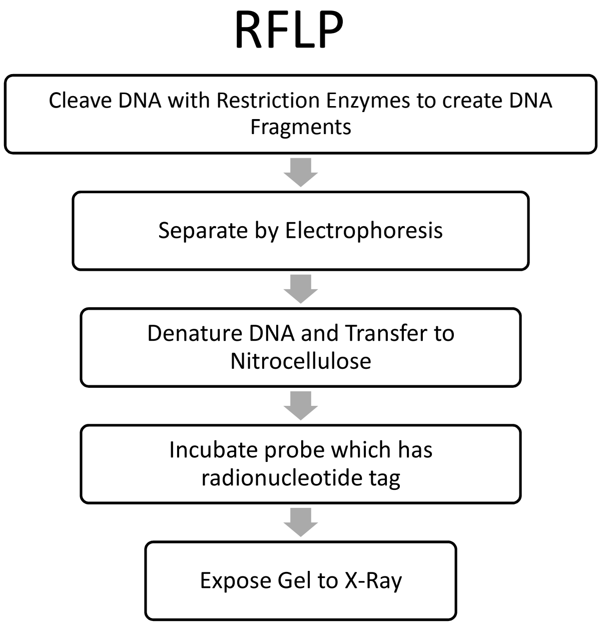
3.2 Procedures

We will now consider several procedures for the collection of genetically related data. Some use the basics of PCR and some do not. The AFLP approach which seems currently best for comparing species relies heavily upon PCR. The approaches we consider are as in the following Table.

Criterion	AFLP	RAPD	Microsatellite SSR	RFLP	Allozymes
Quantity of information	High	High	High	Low	Low
Replicability	High	Variable	High	High	High
Resolution of genetic differences	High	Moderate	High	High	Moderate
Ease of use and development	Moderate	Easy	Difficult	Difficult	Easy

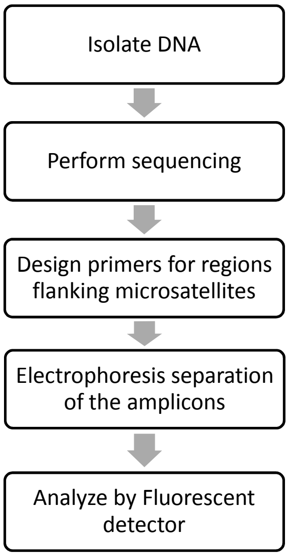
3.2.1 RFLP

Restriction Fragment Length Polymorphisms or RFLP is one of the older mechanisms to obtain DNA fragments to analyze. The approach is detailed in the following Figure. Simply we use restriction enzymes then separate and use a probe to bind to the ends and then use an X ray which can detect the probe areas.


RFLP
Cleave DNA with Restriction Enzymes to create DNA Fragments
Separate by Electrophoresis
Denature DNA and Transfer to Nitrocellulose
Incubate probe which has radionucleotide tag
Expose Gel to X-Ray

3.2.2 Microsatellite

Microsatellites are similar to the RFLP and instead of X-Rays we use fluorescent scans. The details of this approach are shown below. This is a small sequence approach of about six base pairs. Primers and PCR can be applied. The details are shown below.

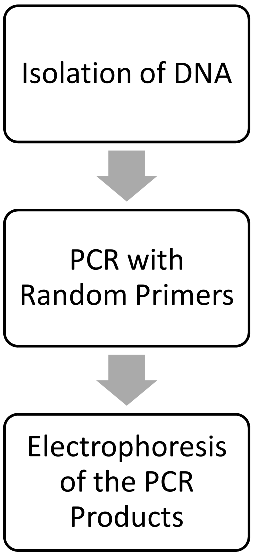


Microsatellite

Isolate DNA
Perform sequencing
Design primers for regions flanking microsatellites
Electrophoresis separation of the amplicons
Analyze by Fluorescent detector

3.2.3 RAPD

RAPD is Random Amplification of Polymorphic DNA. To some degree this is a "shot in the dark" approach. It generates many segments in a random fashion and then one can compare one species or plant to another. The mechanism is shown below in the Figure. This approach has not been used greatly in Hemerocallis analysis.



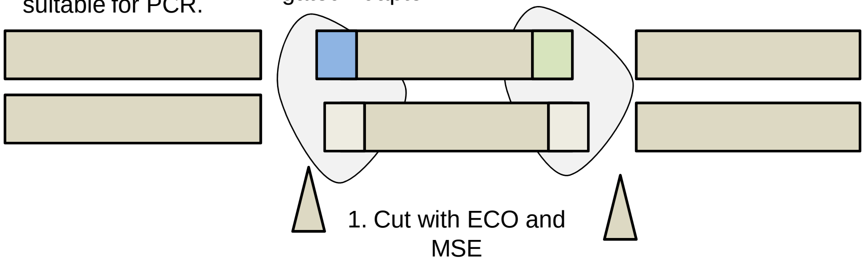
RAPD

3.2.4 AFLP

AFLP or Amplified fragment length polymorphism is an intriguing approach which combines the best of all the other schemes. t gets long fragments, it has the ability to obtain quite a few and it has markers which give good results. It also uses PCR very effectively. The approach is shown in the following steps.

First we take DNA which we have extracted from the cell and then cut it with enzymes and after the cutting we ligate to the ends marker strips which we use to facilitate subsequent PCR.

AFLP I



2. Ligate at each ends adaptors suitable for PCR.

2.1 ECOR1 Ligated Adaptor

2.2 MSE1 Ligated Adaptor

1. Cut with ECO and MSE

Step 1: Take double stranded DNA and first, cut it with two enzymes, ECOR1 and MSE1. After having cleaved the DNA, then, second, ligate to the cuts the marker strips of DNA which can be used to facilitate PCRing the fragment.

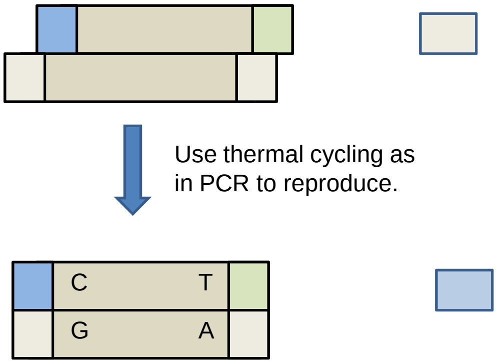
Reference: See *AFLP® Plant Mapping*, Applied Biosystems, 2007,

Second. we then use PCR to effect the growth of many small segments, typically many nucleotides long, and we can create a large amount of these segments. This is shown in the following Figure. This method is what is provided by Applied Biosystems.

AFLP II

Start with some pre-selection using the ligated primaries. Note that:

1. MseI the complementary primer has a 3´ C.
2. EcoRI the complementary primer has a 3´ A or no base addition.
3. PCR provided a preference in multiplication on the two end primed segments.
4. This process acts to purify the batch of segments to those with the two bases ligated.



Use thermal cycling as in PCR to reproduce.

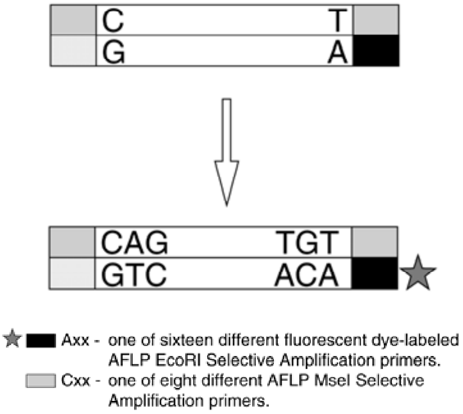
▭ MSE1 adaptor and recognition site and A OR ECOR1 adaptor or recognition site

▭ MSE1 adaptor and recognition site and C

Third, by using tagged primers on the ends of the fragments, we can use the 24 possible sets of primers obtain quite a large and diverse set of cuts. We do this both within a species and between a set of species.

AFLP III

This is a selective PCR process using tagged primers. The primer may be dye labelled and allows for selective processing. Additional PCR amplifications are run to further reduce the complexity of the mixture so that it can be resolved on a polyacrylamide gel. These amplifications use primers chosen from the 24 available AFLP Selective Primers (eight MseI and sixteen EcoRI primers). After PCR amplification with these primers, a portion of each sample is analyzed



★ ■ Axx - one of sixteen different fluorescent dye-labeled AFLP EcoRI Selective Amplification primers.
□ Cxx - one of eight different AFLP MseI Selective Amplification primers.

Then we take the results and create electrophoresis results. This is shown below. In this Figure we show as rows bands of separate fragments which would result from performing an electrophoresis. Each column is a set of bands from a separate species. This simplified diagram shows how we can take many such fragments, from the possible 24 primers and the fact that each enzyme make cuts at different places, we get many possible fragments per plant. In the Figure the dark bands represent a fragment as it may possibly appear in an electrophoresis result. In addition the fragments are longer in the number of nucleotides so we can get a finer set of resolution than we could possibly attain in a single RFLP or RSS.

In addition, we can now use this data to create a set of relationships. Movement of bands means changes in genes, specifically nucleotides. Thus for small change we get a close match and for large changes we may get many splits. We then will use this data to create what we call a distance matrix which is a set of measure for showing how different species vary at the genetic level.

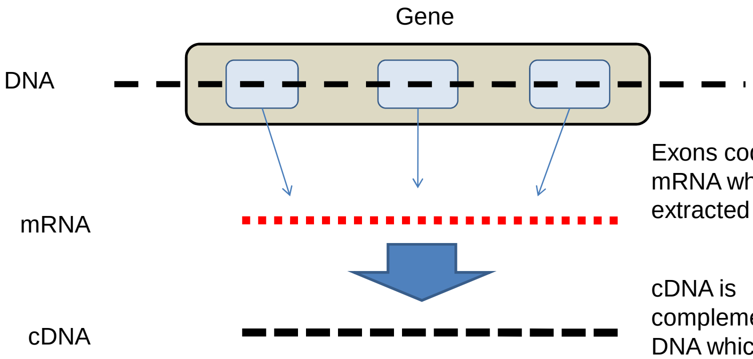
Band	Aurantiaca	Altissima	Dumortieri	Middendorfii	Fulva	Flava	Hakuuensis	Thunbergii	Minor
1	■		■	■	■		■	■	■
2		■	■						
3				■	■	■	■	■	■
4	■	■							
5	■	■	■	■		■	■	■	■
6									
7			■	■	■	■	■		
8									
9		■	■	■					
10									
11									
12	■	■	■			■	■		
13									
14					■	■	■	■	■
15									
16									
17		■	■	■					
18									
19						■	■	■	
20									
21									
22	■	■	■	■	■	■	■	■	■
23									

3.2.5 Microarrays

Microarrays is a unique approach which allows for the analysis of millions of samples, it is a marriage of high tech solid state chip technology with DNA bonding. We describe it in the following four steps, each accompanied by a Figure.

Step 1: The first step in a micro array is the production of cDNA, or complementary DNA. cDNA is that set of nucleotides which account for the encoding of mRNA. It does not include the non-coding regions which are the introns.

Microarray I



Exons code for mRNA which is extracted from cell.

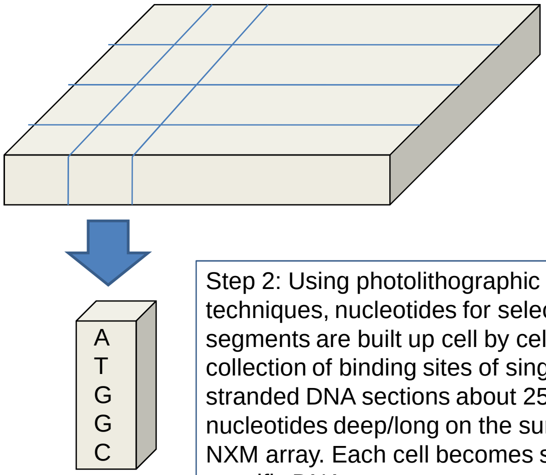
cDNA is complementary DNA which is DNA regenerated from mRNA and it is same as original gene but does not have Introns.

Step 1: from the desired DNA to be analyzed the mRNA is extracted and used to generate specific cDNA.

Step 2: In a separate environment we make the microcell. This is created in a manner identical to the making of integrated circuits which entails photo-masking techniques. Instead of silicon we used nucleotides. The array has millions of small holes in an array like manner. Each hole we fill with nucleotide, one nucleotide at a time.

For example, we can use the columns to drop DNA from each species sample and we then use each row with a set of 25 probe nucleotides to determine if that matching gene is present. The rows may be entered to match known genes, and using 25 sequential nucleotides we can fairly accurately get a gene. There are 4^{25} possible sequences and in Hemerocallis there are a few thousand genes, and we must know them otherwise we would be just "shooting in the dark". Microarrays do require knowledge of the CDNA library at least of key genes. We know, for example, from the work of Mol and Winkel Shirley the genes that control the secondary pathways for color. This we have discussed elsewhere.

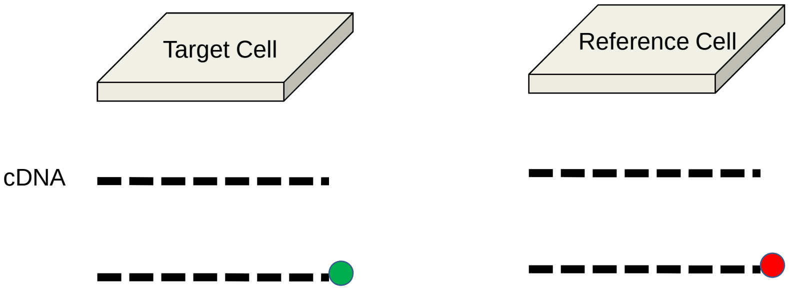
Microarray II



Step 2: Using photolithographic techniques, nucleotides for selected cDNA segments are built up cell by cell creating a collection of binding sites of single stranded DNA sections about 25 nucleotides deep/long on the surface of an NXM array. Each cell becomes sticky for a specific DNA segment.

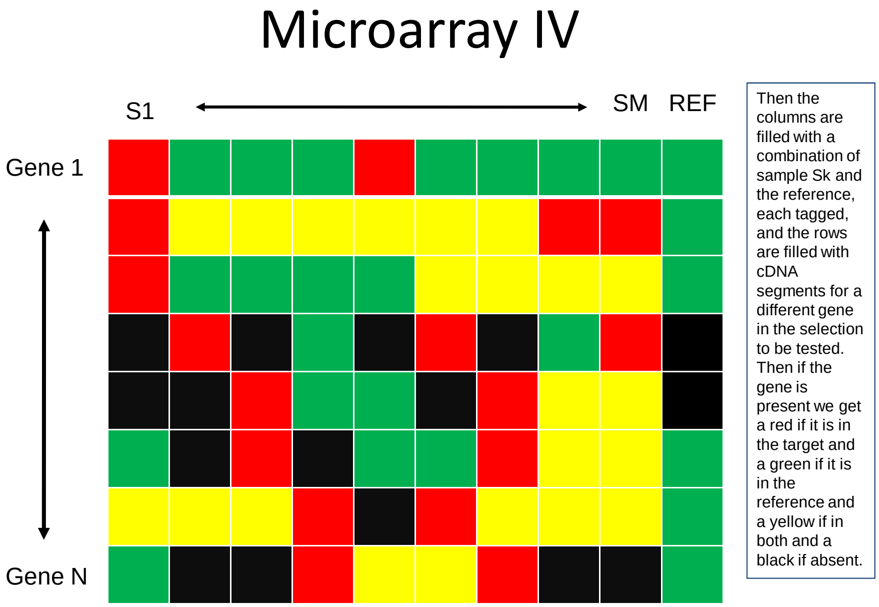
Step 3: Now we take two DNA samples, one from what we call the Target, the plane we wish to categorize, and we use a reference plat as well, say H fulva. We then take the segments we collected in step one and tag then with green or red tags, green say for the Target and Red for the Reference.

Microarray III



Step 3: For the DNA to be analyzed and a "Reference" target DNA, the mRNA is extracted from each and the cDNA is produced for every gene in the cells to be analyzed, and then it is tagged with a dye which is red for one and green for the other. Typically we tag the target red and Reference green.

Step 4: We then take the samples from the differing plants, one in each column, and look at the array. If the microarray cell has the gene sequence we are seeking to march, and the Target has that sequence, it will bond and stick. If the Reference has it, it too will bond. If we just get the Target the cell will be green, if we just get the Reference the cell is red, if we get both the cell turns yellow, and if we have neither the cell is black. The result of a sample scan is shown below.



Microarray IV

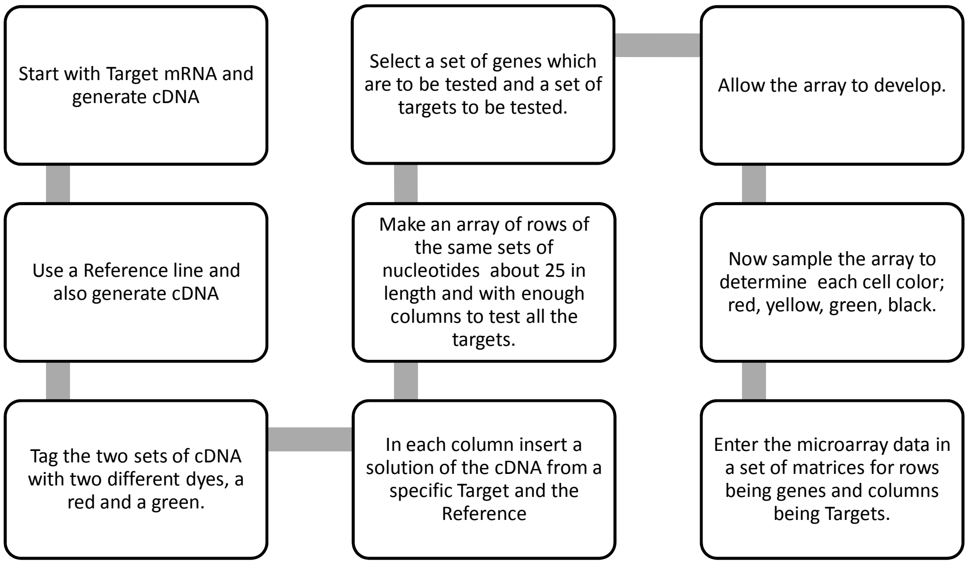
	S1								SM	REF
Gene 1	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩
	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟩
	🟥	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟩
	⬛	🟥	⬛	🟩	⬛	🟥	⬛	🟩	🟥	⬛
	⬛	⬛	🟥	🟩	🟩	⬛	🟥	🟨	🟨	⬛
	🟩	⬛	🟥	⬛	🟩	🟩	🟥	🟨	🟨	🟩
	🟨	🟨	🟨	🟥	⬛	🟥	🟨	🟨	🟨	🟩
Gene N	🟩	⬛	⬛	🟥	🟨	🟨	🟥	⬛	⬛	🟩

Then the columns are filled with a combination of sample Sk and the reference, each tagged, and the rows are filled with cDNA segments for a different gene in the selection to be tested. Then if the gene is present we get a red if it is in the target and a green if it is in the reference and a yellow if in both and a black if absent.

Now, we even get to try and look at the intensity of the red, green, or yellow. This we can try to see how much is expressed not just whether it is or is not. We will not discuss that here. In the above matrix we can see that many genes are expressed in one or both or none. If we have enough genes than we can argue we have the basis for an exceptionally good means to develop a classification.

In the following Figure we summarize the microarray process.

Microarray Summary



Start with Target mRNA and generate cDNA

Use a Reference line and also generate cDNA

Tag the two sets of cDNA with two different dyes, a red and a green.

In each column insert a solution of the cDNA from a specific Target and the Reference

Make an array of rows of the same sets of nucleotides about 25 in length and with enough columns to test all the targets.

Select a set of genes which are to be tested and a set of targets to be tested.

Allow the array to develop.

Now sample the array to determine each cell color; red, yellow, green, black.

Enter the microarray data in a set of matrices for rows being genes and columns being Targets.

3.3 Comparisons

We can now compare the various methods we believe are effective. The three are the AFLP method, the second if the microarray and the third is total gene mapping. We defer the latter for the present.

	AFLP	Microarray	DNA
Advantage	Fast Can use many markers Can use NJ technique	Uses specific targeted genes Can provide for genetic variation with some time evolutionary analysis Can use NJ technique	Uses actual nucleotide sequences Can be used to determine time of evolution
Disadvantage	Limited number markers Does not reflect true genetic comparison Sequences are generally targets of opportunity	Requires known Genes	Requires large data sets Costly Analysis is complicated and should use ML techniques

4 CLASSIFICATION TECHNIQUES

We now discuss various classification techniques which use as input results from some form of DNA analysis such as the methods we have just discussed. Our goal is:

- Develop tools which can project relationships from data obtained using genetic material.
- Relate separate species to one another in a definable and metric based format.
- Look for consistency between gene based relationships and species based relationships.

4.1 Principles

Trees are a graphical manner to represent relationships. The specific relationship we may wish to represent is one that reflects evolutionary relationships in time, namely which came first and which came after. In the development of trees using morphology we may look at sessile versus branched as a factor which may reflect temporal evolution. Namely in the monocots the sessile character may have some reason based upon paleobotany to have preceded the branching character. Thus, if we had such a basis or justification we would try to incorporate that factor.

The basic principles we try to use in developing trees are:

1. Parsimony: This is the Ockham's razor principle of using the simplest answer.

2. Bifurcation: New species come out one at a time because the enabling genetic change is one gene or one nucleotide at a time.

3. Time is reflected in a Distance Measure: There can always be create a distance measure between species based upon some set of characteristics. This measure may be morphological as having sessile versus branched, the length of a petal, the color of the flower, the width of a leaf. Or it may be genetic, the presence of an enzyme, the presence or absence of a specific sequence of nucleotides, or the number of restriction fragments across all chromosomes. The distance measure takes the difference and maps them to a number. The number then is related to them, namely how long does it take for a nucleotide to change.

4. Trees have inherent structures and the Classification meets the structure of the tree more than it may meet reality: We use the theory of trees to develop the dendrograms. This may limit what really happens.

5. Disbelief in "Black Swans": Black Swans are unexpected events. They may be a catastrophe, a crisis never expected, and some upheaval in nature. The models we generally develop assume the past acts continuously and that change is not the result of some dramatic change.

These principles make what we do in classification easier. We can model small and understandable change. We cannot model the unpredictable. Thus we should be aware that these models are the result of these assumptions and perhaps even more yet to be articulated. The classic systematics practitioner rationalizes what they do. One need look at the texts by Judd or those by Felsenstein, brilliant efforts and in many ways documents which reflect a reality. Yet they all lack the Black Swan which we know hides just in the shadows of all reality.

4.2 Measurements and Metrics

The first step is to define and develop metrics, distance reflective of change, measures that map measurements of reality into manipulateable numbers. In our analysis we will focus on data measurements resulting from the techniques we detailed in the previous section. As such we can look at three general areas:

1. Gene Dynamics (Nucleotide Changes of ATGC): This may use the classic Jukes Cantor measure of change of nucleotides which assumes equal probability of nucleotide change per unit time. We may measure the nucleotide strings, nucleotide by nucleotide and from this try to see how they may be best arranged so that we may characterize evolution consistent with a model of change based upon some reality.

2. Inferred Genetic Distances: This approach uses data such as AFLP data and the like and then defines a distance between them in some manner which reflects gene change. In our case we use 1, 0 as binary. Could also use measure of number of nucleotide changes if that could be determined. Microarray data could be used here as well.

3. Non Genetic based upon ODU. These may also be clustering techniques and it does not utilize measurements of the type we look at here.

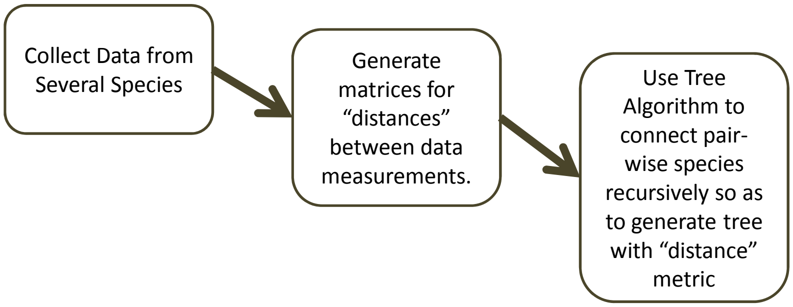
The overall process which we are to follow is depicted in the following figure. It is a simple three step process:

1. Obtain the raw genetic data.

2. Create distance measure based upon the data.

3. Employ the distance measure to create trees.

Process



Collect Data from Several Species

Generate matrices for “distances” between data measurements.

Use Tree Algorithm to connect pair-wise species recursively so as to generate tree with “distance” metric

The following is a repeat of the AFLP data that we may collect in an experiment. It is a rendition of an electrophoresis plot of the AFLP marker sequences. We take this chart can convert it to a distance matrix.

AFLP Data

Band	Aurantiaca	Altissima	Dumortieri	Middendorfii	Fulva	Flava	Hakuuensis	Thunbergii	Minor
1	■		■	■	■		■	■	■
2		■	■						
3				■	■	■	■	■	■
4	■	■							
5	■	■	■	■		■	■	■	■
6									
7			■	■	■	■	■		
8									
9		■	■	■					
10									
11									
12	■	■	■			■	■		
13									
14					■	■	■	■	■
15									
16									
17		■	■	■					
18									
19						■	■	■	
20									
21									
22	■	■	■	■	■	■	■	■	■
23									

The following depicts a distance matrix for this AFLP data. Let us assume we have M species and N measurements for each species in the electrophoresis chart. Thus the above has M columns and N rows. We then create a matrix which measures the distance between two species. It is as below. We look at the two columns and we generate a distance as the number of electrophoresis bands where they differ.

We define the distance measure as follows:

$$D_{i,j} = \frac{1}{N}\sum_{k=1}^{N} d_{i,j}$$

$$d_{i,j} = \begin{cases} 1 \textit{ if there is a band in one and not the other} \\ \textit{0 if there is a band in both or no band in both} \end{cases}$$

The following shows the distance matrix between the species. Again this is a repeat of details we presented earlier. Note that the matrix is symmetric.

Distance Matrix

	Aurantiaca	Altissima	Dumortieri	Middendorfii	Fulva	Flava	Hakuuensis	Thunbergii	Minor
Aurantiaca	0	1	3	4	7	2	9	3	5
Altissima	1	0	4	7	9	2	3	5	6
Dumortieri	3		0	3	6	7	9	3	2
Middendorfii	4			0	2	6	9	3	5
Fulva	7				0	5	2	3	9
Flava	2					0	8	2	5
Hakuunensis	9						0	2	9
Thunbergii	3							0	4
Minor	5								0

There are a few issues we must be concerned with. First are the options of a measure. For example: (i) develop binary measures such as {0,1} values based expression or non-expression of gene and (ii) create artifact distances as a measure expression by measuring the density of the color; thus a variable on the interval [-1,1]. Second there are many issues that need to be focused on such as: (i) Sensitivity of the measurements, (ii) Use of a reference mix and (iii) All issues related to errors in microarrays and their measurements.

4.3 Techniques for Trees

The following are the principle techniques found in the development of Trees:

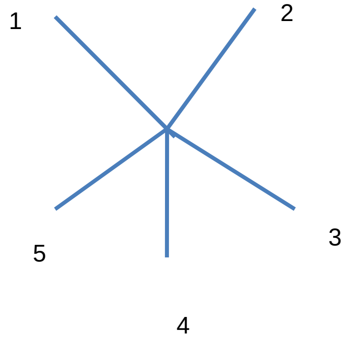
- Neighbor Joining: Tries to get a tree with the best possible fit of an additive rooted binary tree.
- Maximum Likelihood: Assumes an underlying transition process and then attempts to create a tree based upon a best fit to that process.
- Maximum Parsimony
- Generalized Neighbor Joining
- Weighted Neighbor Joining
- Un-weighted Pair Group with Arithmetic Mean (UPGMA)
- Minimum Evolution
- Fitch-Margoliash-Least Squares Fit

We will look at two of these; Neighbor Joining and Maximum Likelihood.

4.3.1 Neighbor Joining

The Neighbor Joining scheme is used frequently. It was developed in the mid 80s and was modified to correct initial errors in the analysis and also to improve the running time of the algorithm.

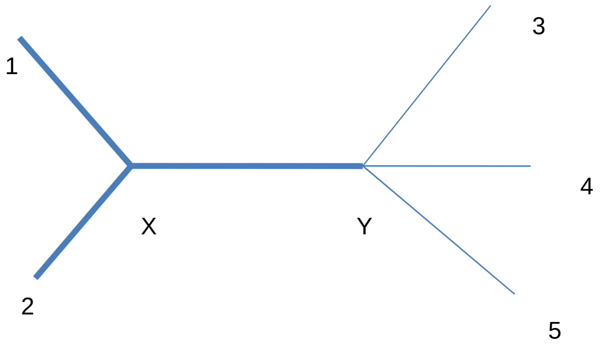
Tree Generation I



Start with say 5 of the possible species. They are all connected and we know we have a distance matrix which give a “distance” between all pairs of this collection. We now want to create a “tree” on a pair-wise basis so that there is some sound relationship between the end points, namely the species.

Then we begin the development of a tree. This we depict below. Inherent in this process is the assumption that species split from a common ancestor in pairs. Namely we have a binary set of nodes; we never get three species at a split, only two.

Tree Generation II



We start the tree process by selecting in some manner pairs of “closest” end points and then building this out.

We focus on Trees which are additive Trees. A tree is a connected graph which has no cycles. In a tree there is a unique path between every pair of vertices. An Additive Tree is a tree which has certain properties. Namely in an additive tree we have:

$$L_{ij} = \textit{length of any path in the tree between any two points.}$$

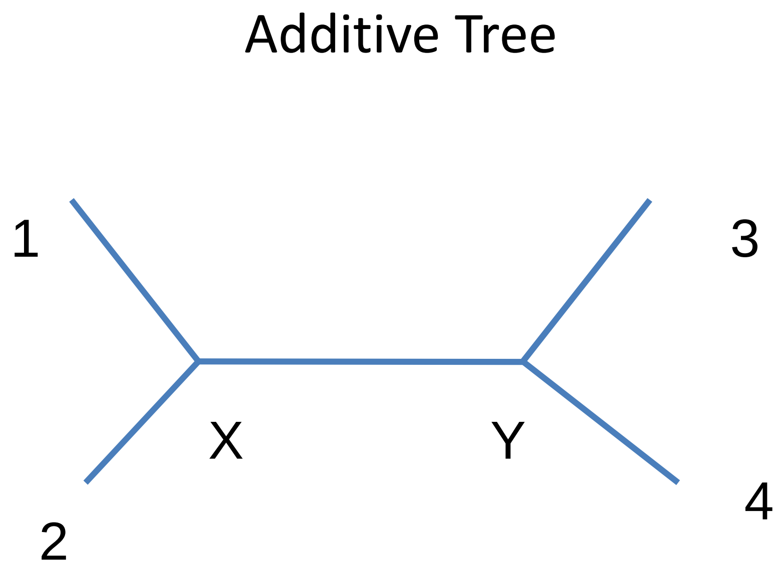
$$L_{ij} = L_{in} + L_{n,n+1} + \ldots + L_{N,j};\ \textit{where each are lengths of contiguous segments}$$

Thus in our example we have:

$$D_{1,2} = L_{1,X} + L_{X,2}$$

A binary tree is a tree with bifurcated ends, namely only two vertices at each branch in a tree. Finally a rooted tree is a tree is a binary tree with a single starting point reflecting evolutionary trends.

Consider the following simple Tree.



We can show that the entries in the distance matrix and the path lengths can be calculated.

$$D_{12} = L_{1X} + L_{X2}$$
$$D_{34} = L_{3Y} + L_{Y4}$$
$$D_{13} = L_{1X} + L_{XY} + L_{Y3}$$
$$D_{23} = L_{2X} + L_{XY} + L_{Y3}$$
$$D_{24} = L_{2X} + L_{XY} + L_{Y4}$$

And we can write this as:

$$\begin{bmatrix} D_{12} \\ D_{13} \\ D_{14} \\ D_{23} \\ D_{24} \\ D_{34} \end{bmatrix} = \begin{bmatrix} 11000 \\ 10101 \\ 10011 \\ 01101 \\ 00110 \end{bmatrix} \begin{bmatrix} L_{1X} \\ L_{2X} \\ L_{3Y} \\ L_{4Y} \\ L_{XY} \end{bmatrix}$$

We can now state the Neighbor Joining Algorithm:

1. We begin with all of the vertices in a star formation and then we compute for each pair of vertices the factor S_{ij}. We select the pair with the least value. Recall:

$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^{N} L_{iY}$$

2. Using the relationships between the L and D elements we can write this as:

$$S_{12} = \frac{1}{2(N-2)} \sum_{i=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{j=3}^{N} \sum_{3 \le i}^{j-1} D_{ij}$$

3. Calculate this for all pairs and select the pair to join which has the smallest S value.

4. Now we have a tree with N-1 vertices But the new vertex is the combination of 1 and 2, we call it X. We now need to obtain the new D values that relate to this new vertex. We define that as:

$$D_{xj} = \frac{\left(D_{1j} + D_{2j}\right)}{2}; (3 \le j \le N)$$

5. We then go back to step 2 and use these new values and select again the new pair that gives the smallest S value. We repeat this process until we have all pairs.

6. The dendrogram is the result using NJ and AFLP data for the various Hemerocallis species.

4.3.2 Maximum Likelihood

Maximum Likelihood is an approach to classification using genetic data, genes specifically, and it incorporates details about the changes in the genes over time. The maximum likelihood approach assumes that we have obtained a mapping of the gene or some gene segment down to the nucleotide. Then it assumes we have the same segments for the other species we wish to compare. Let us assume we have twelve species and we have the following twelve 25 nucleotide long segments. We can assume that they come from a cDNA, recalling that cDNA is made from mRNA using a reverse transcriptase. Thus we have the following as in the Table:

Species	*cDNA Segment*
altissima	**AATTCTACTTACTTACTGGACCAGT**
aurantiaca	**AATTCGGCTTGCGTACTGGACCAGT**
citrina	**AATTCCCCTTACTTACTGGACCAGT**
coreana	**AATTCGGCTTACGCGCTGGACCAGT**
dumortierii	**AATTCGGCTTACTTACTGGACCTAA**
flava	**AACGCGGCTTACTTACTGGACCAGT**
fulva	**AATTCGGCTTTAATACTGGACCAGT**
hakunensis	**AATTCGGCGGACTTACTGGACCAGT**
middendorfii	**AATTCGGCTTACTTACTCCACCAGT**
minor	**AATTCGGCAAACTTACTGGACCAGT**
multiflora	**CCTTCGGCTTACTTACTGGACCAGT**
thunbergii	**AATTCGGCTTACGGACTGGACCAGT**

We may hypothesize that the original sequence is **AATTCGGCTTACTTACTGGACCAGT.** If we did then we have noted the changes in the sequences by the red nucleotides in each of them. We then pose the following problem:

1. Assume we have 12 nucleotide sequences from a segment of cDNA we know to be a useful segment in determining a plant characteristic, such as color.

2. For each of the segments, that is, for each species, define a vector of dimension 25X1 as z(n) for each of the 12 species.

3. Assume the following:

a) There existed a common ancestor for all of these species.

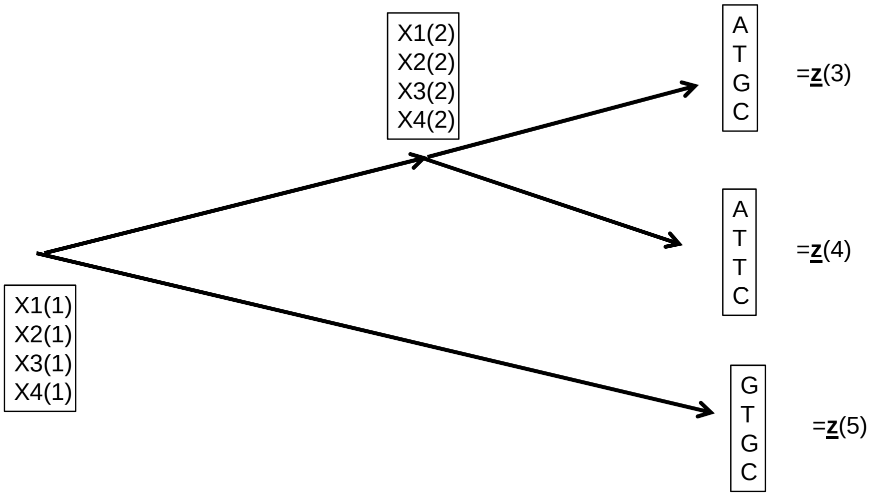
b) Evolution occurs at one nucleotide change at a time and is binary. Namely we do not get multiple nucleotide changes and we do not get binary change happening simultaneously.

c) Assume that we can ascribe a probability to a single nucleotide change, and we may or may not know the value and the value may or may not remain constant over the time horizon.

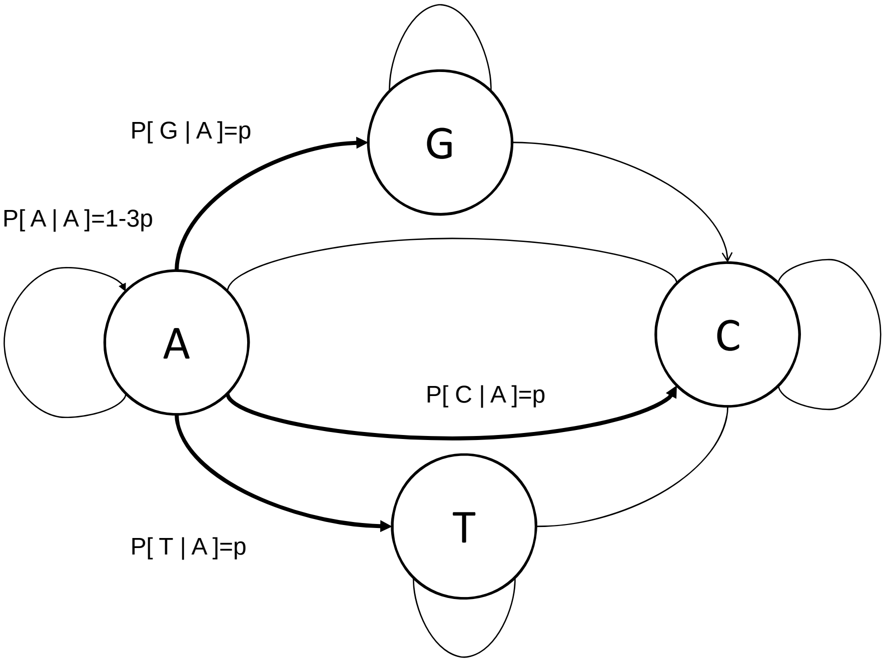
Thus with these assumptions there will exist a single rooted tree for this set of species.

4. The changes that occur do so independently. That is we have a Markov process.

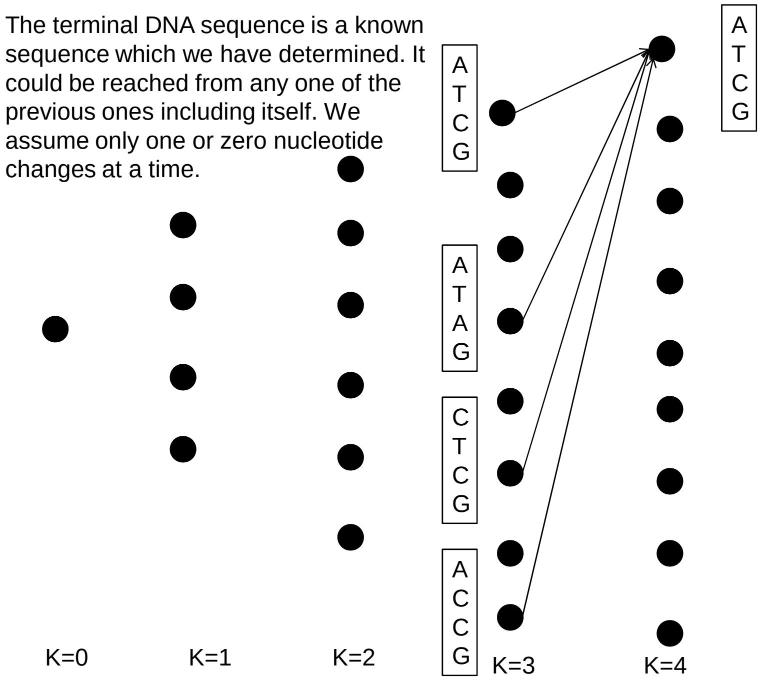
The following Figure depicts what we are posing. The internal nodes, assumed to be 25 nucleotide sequences also are labelled as x(n). They are 25X1 vectors as well



We can model the nucleotide changes as a Markov process and we can use a finite state machine to do so. This we show below. The probability of changing a nucleotide is p and is the same for all changes. This is the simplest model possible. One may look at various other and more complex models.

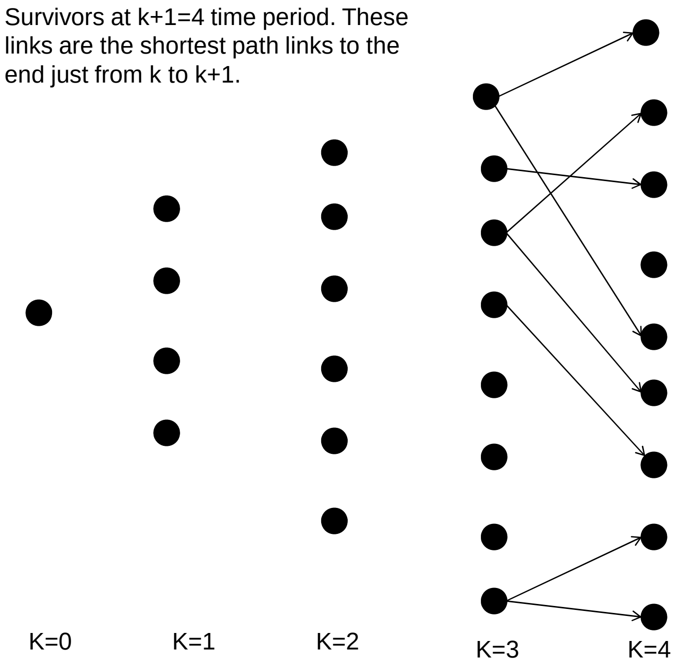


The problem can now be further posed as shown in the Figure below. We have the end points on a sequence of changes. We know the limits we have placed on the changes and we now want to find a process which will give us the "best" set of past changes so that we get what we observe in the 12 different 25 nucleotide sequences. In addition we want to get a single end point tree which is generated by single bifurcations. The following Figure looks at the end point.



The terminal DNA sequence is a known sequence which we have determined. It could be reached from any one of the previous ones including itself. We assume only one or zero nucleotide changes at a time.

if we just look at the last steps, we know that if we had some algorithm which gave us the best path then there would be some best path to every one of the know end elements. Then we would ask how we got to them. This end element best path is shown below.



Survivors at k+1=4 time period. These links are the shortest path links to the end just from k to k+1.

Now the principles of the maximum likelihood approach are:

- Deals with DNA Sequences
- Known rates of nucleotide change per unit time
- Changes result in two new paths and no more at any one time
- Changes always go "forward", no crossing or reverses

Further

- Assume we have ACTG type nucleotides
- Assume that there is a rate of change of α per unit time. Thus over T units of time the probability of a single nucleotide change is p which is αT
- Assume all are equally the same
- Then we have a finite state machine model for the change

The problem is then to find the sequence of change states which lead to the known final states so that the sequence maximizes the a posteriori probability or as in the following:

$$\max\ p(x(1)...x(n)|z(1)...z(m)) = \max \frac{p(z|x)p(x)}{p(z)}$$

This is the maximum likelihood approach. Let us explain it a bit.

1. The probability density, p(x|z) is the a posteriori probability of some or all of the internal nodes, we call them x, give the observed end nodes, and we call them z.

2. We want to find the set of all possible internal nodes, the set of all possible xs, that can yield the observed z, and we want that specific set of x which maximizes the a posteriori probability. Well one may ask why that is a good thing to do. There have been many analyses of this problem but the best approach is looking at detection of targets in radar, where this was most effectively used. Selecting this point maximizes the target hit probability and minimizes the false alarm rates.

But we can also write the above in terms of the p(z|x) and then the p(x). We can reject the p(z) since it has no impact on choosing the x.

Now since we have structured this with Markov processes, and since this means that changes depend only on their immediate past we can write:

$$p(z\,|\,x) = p(z(1)\,|\,x(k))\,p(z(2)\,|\,x(j))...p(z(m)\,|\,x(r))$$

That is we can write the $p(z\,|\,x) = p(z(1)\,|\,x(k))\,p(z(2)\,|\,x(j))...p(z(m)\,|\,x(r))$

And recall that we have:

$$p(x(j)\,|\,x(k)) = \begin{cases} 1-3p \\ p \end{cases}$$

Thus in our earlier initial map with end nodes we can write for each the following:

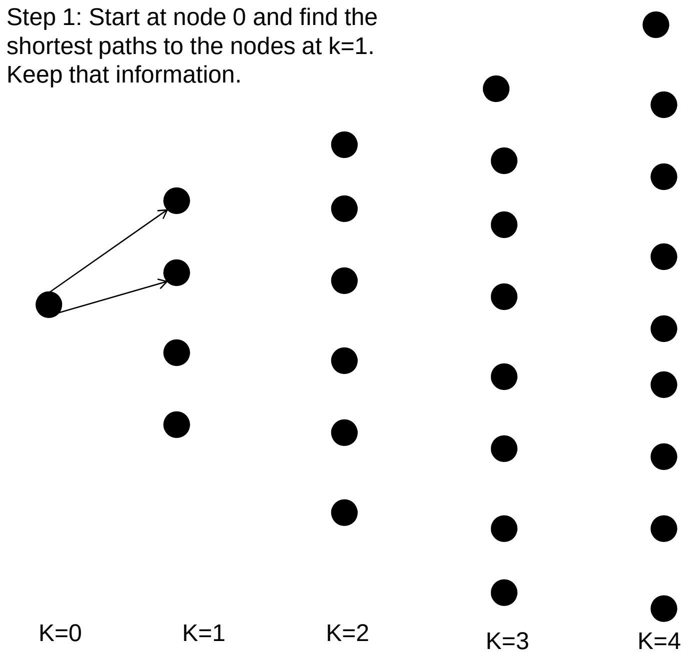
$$p(z(3)\,|\,x(2),x(1)) = p(z(3)\,|\,x(2))\,p(x(2)\,|\,x(1))$$
$$p(z(4)\,|\,x(2),x(1)) = p(z(4)\,|\,x(2))\,p(x(2)\,|\,x(1))$$
$$p(z(5)\,|\,x(1)) = p(z(5)\,|\,x(1))$$

We find it more convenient to define a distance value defined as:

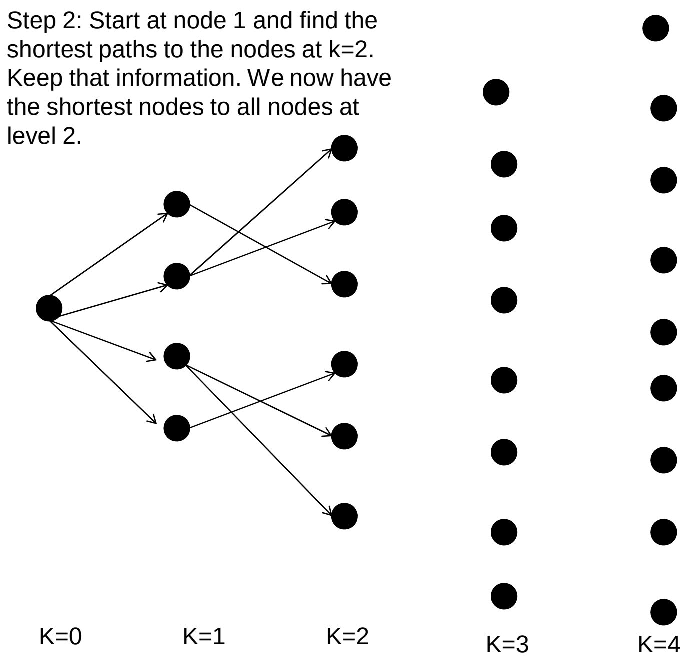
$$\lambda(\zeta(k)) = -\ln p(x(k+1)\,|\,x(k)) - \ln p(z(k)\,|\,x(k))$$

Thus instead of maximizing the probability we minimize the distance as defined above. In addition we can perform computations better this way. Thus we are seeking the minimum length path through the network where we define length as above. This is called the Viterbi algorithm and was used first in decoding convolutional codes.

The algorithm is developed graphically as follows:

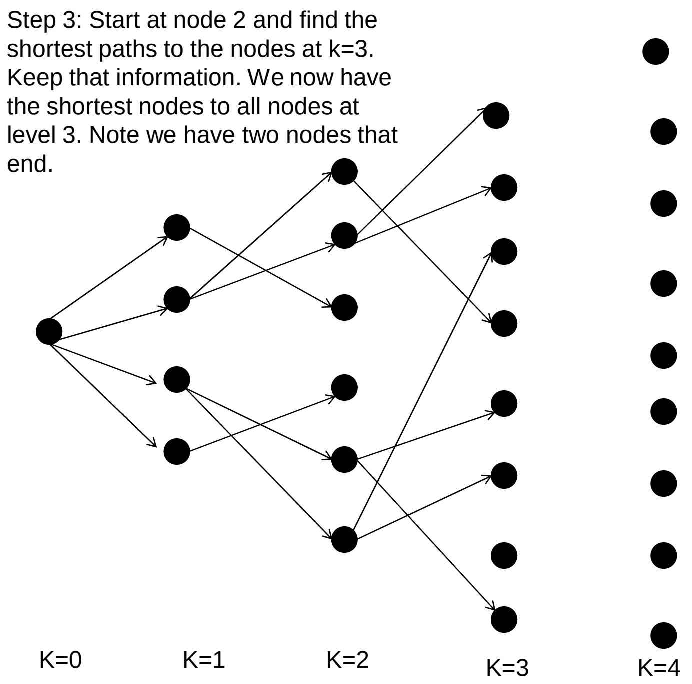

Step 1: Start at node 0 and find the shortest paths to the nodes at k=1. Keep that information.
K=0
K=1
K=2
K=3
K=4

Thus step 1 starts at the beginning. Here we assume the beginning for some node. We can and will do it for all possible nodes. Recall that for a 25 nucleotide sequence we have 4^{25} nodes. Then we go to step 2 as below.



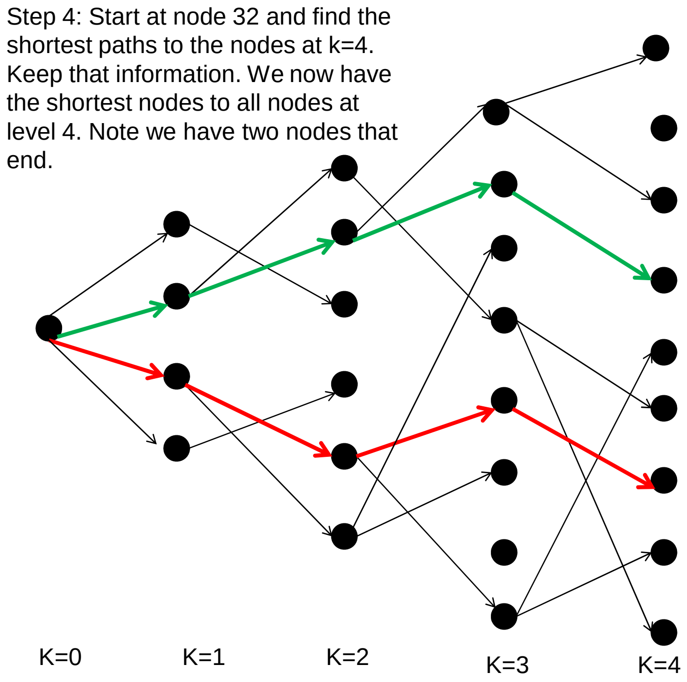
Step 2: Start at node 1 and find the shortest paths to the nodes at k=2. Keep that information. We now have the shortest nodes to all nodes at level 2.

In the above step we start culling out nodes. We keep only at k-2 those paths of least length where we define length as above. We continue as below:



Step 3: Start at node 2 and find the shortest paths to the nodes at k=3. Keep that information. We now have the shortest nodes to all nodes at level 3. Note we have two nodes that end.

Finally we get to the ends, and each path to each end node is the least distance path.


Step 4: Start at node 32 and find the shortest paths to the nodes at k=4. Keep that information. We now have the shortest nodes to all nodes at level 4. Note we have two nodes that end.
K=0
K=1
K=2
K=3
K=4

Note several things about this algorithm.

1. All paths are minimum length from the selected initial point. If we change the initial point we get a whole new set of paths. To determine the best initial point we do this for all possible initial points and choose the one with the least sum of the lengths. This is computationally intense.

2. The red and green paths above show details of two specific paths. Note in both there is no branching at step k=3. The other end points branch at k=3. This we have a binary tree perforce of the assumptions and not a result of anything we see in the data. The assumptions are control the end result often more than the data so beware assumptions.

3. The resulting tree becomes evident. One need just follow the path.

Many authors have tried to explain this approach to no avail. I have seen such works as that of Durbin et al which make it totally incomprehensible! One should beware those who have notation which is incomprehensible.

5 APPLICATION TO HEMEROCALLIS

The first extensive efforts at taxonomy within the daylily were attempted by A.B. Stout (1934), in which two major classifications were proposed: those having branched scapes (Euhemera) and those without branched scapes (Dihemera). Stout's classification, however, is now not generally well accepted. A more recent classification of daylily species into five major groups is presented by Erhardt (1992), and generally supported by the AFLP data in the present study. Erhardt's classification of the five groups comprises (1) *fulva, (2) citrina, (3) middendorffii, (4) nana, and* (5) *multiflora.*

Utilizing neighbor-joining analysis, the six *H. fulvas* were distinctly separated from the other species. Clustering within the *fulvas* also supported some fine-scale taxonomic classifications. For example, the distinction described by Erhardt between the two fulva double-flowered genotypes ‘Kwanso’ and ‘Flore Pleno’ is reflected in the molecular data. Within the *middendorffii* group, *H. dumortierri*, *Hemerocallis middendorffii* and *Hemerocallis hakunensis* all grouped together as proposed by Erhardt.

However, the distinction between the *citrina* group and the *middendorffii* group was not well defined and contained some overlap. *H. citrina* and *Hemerocallis minor* were grouped together as proposed by Erhardt, but were also grouped with members of the *middendorffii* group. Erhardt had proposed a close relationship between two other members of the citrina group, *Hemerocallis lilioasphodelus* and *Hemerocallis thunbergii*, which was well supported by our data, but they did not closely group with the other *citrina* members. In fact, our data suggest that the *middendorffii* group and the *citrina* group should be merged into one large taxonomic group.

The only major anomalies among the species analysis were supposed clonal variants of *H. citrina* (var. Vespertina) and *Hemerocallis dumortierii* (var. Sieboldii). While both did cluster within the *middendorffii-citrina* group, they did not closely group with their respective parental clones from which they were supposedly derived. Traditionally, there have been a number of variants of *H. dumortierrii* in commerce.

Thus, the variety Sieboldii may or may not include the traditional species *H. dumortierrii* as a direct ancestor even though there are phenotypic similarities. *H. citrina* is self-incompatible and thus any variant arising from it would have to be obtained from an outcross. Hence, these genotypes may either have arisen via cross-pollination or may represent distinctly different genotypes. The following Table recounts Tompkins et al AFLP data.

Tomkins and his team performed analyses on the dozens species and hybrids, a massive number but readily doable with AFLP. The following Table depicts the targeted species and the year they were identified.

Genotype	***Year***
H. citrina	1897
H. citrina var. Vespertina	1941
H. dumortierii	1830
H. dumortierii var. Sieboldii	Unknown
H. fulva Europa	1762
H. fulva Flore Pleno	1860
H. fulva var Kwanso	1860
H. fulva var Maculata	1895
H. fulva var Rosea	1924
H. fulva var Sempervirens	1966
H. hakunensis	1943
H. lilioasphodelus	1576
H. middendorffii	1860
H. minor	1748
H. thunbergii	1873

The AFLP primers used on these species and the hybrids are shown in the following Table.

Primer combination	*Total number of bands*	*Polymorhic bands*	*Polymorhism (%)*	*Scored bands*
E-AAG/M-CAA	126	93	74	None
E-AAG/M-CAC	135	109	81	None
E-ACC/M-CAA	130	109	84	None
E-ACC/M-CAC	103	84	82	61
E-ACC/M-CAG	87	66	76	36
E-ACT/M-CAT	136	108	79	None
E-ACT/M-CTT	107	84	78	55
E-ACT/M-CTA	82	63	77	None
	Total 906	Total 716	Mean 79	Total 152

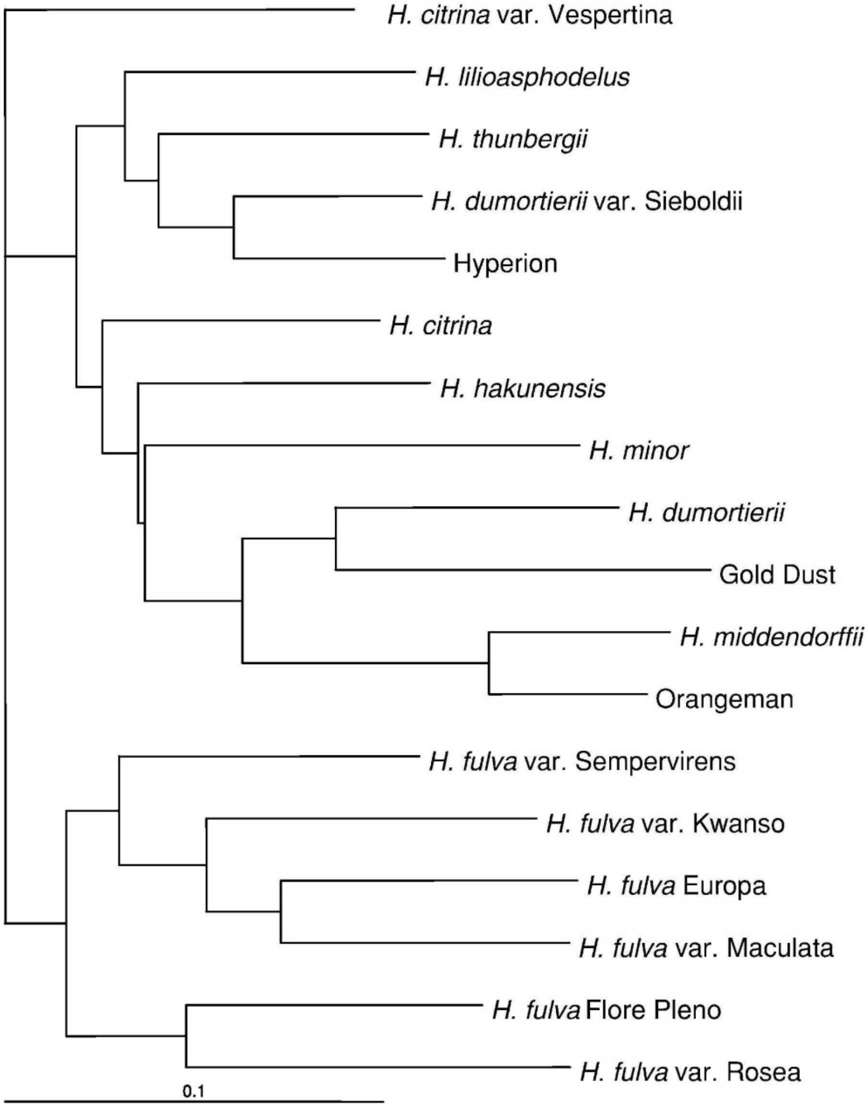
Using this data and employing the NJ technique, Tomkins et al have obtained the following dendrogram. It is effectively a classification using the AFLP markers. The following observations can be made:

1. The species identified as H citrina Vespertina is ranked as the initial root species.

2. H fulva (Europa) the triploid variety is shown to be an offshoot from the split with H fulva kwanso.

3. H citrina as species is an offshoot of Vespertina and is aligned with minor, hakunensis and dumortieri. This seems a bit strange.

4. The H fulvas are all grouped together.



Tomkins summarizes his paper as follows:

"Of particular interest are genetic relationships among species and early cultivars to determine if taxonomic classifications originally performed based on phenotype would be confirmed by molecular relationships obtained in the present study. Therefore, neighbor-joining analysis was carried out on the species and the early cultivars group. The resulting dendrogram is shown…. Taxonomy in the daylily has undergone recent changes and is still somewhat open to conjecture. For our purposes, the AFLP data will be discussed in the context of recent classifications described by Erhardt (1992). Taxonomic classifications were generally supported by the AFLP data. The six H. fulva species all clustered together separately from the other species, which formed a separate cluster and were generally grouped according to Erhardt's proposed group

classifications for the other species. Within this group fell the three early cultivars which showed close relationships to their respective ancestral species progenitors as described in the 1893 to 1957 Hemerocallis checklist... The only anomalies were two clonal variants of Hemerocallis citrina (var. Vespertina) and Hemerocallis dumortieri (var. Sieboldii)."

He then continues:

"Utilizing neighbor-joining analysis, the six H. fulvas were distinctly separated from the other species. Clustering within the fulvas also supported some fine-scale taxonomic classifications. For example, the distinction described by Erhardt between the two fulva double-flowered genotypes 'Kwanso' and 'Flore Pleno' is reflected in the molecular data. Within the middendorffii group, H. dumortierri, Hemerocallis middendorffii and Hemerocallis hakunensis all grouped together as proposed by Erhardt. However, the distinction between the citrina group and the middendorffii group was not well defined and contained some overlap. H. citrina and Hemerocallis minor were grouped together as proposed by Erhardt, but were also grouped with members of the middendorffii group. Erhardt had proposed a close relationship between two other members of the citrina group, Hemerocallis lilioasphodelus and Hemerocallis thunbergii, which was well supported by our data, but they did not closely group with the other citrina members. In fact, our data suggest that the middendorffii group and the citrina group should be merged into one large taxonomic group"

If one reads Tomkins carefully, there is a great deal of ambiguity present. He seems to be trying to keep with Erhardt but he continually diverges. Thus there are still open issues as to Classification using this data.

6 CONCLUSIONS

The purpose of this paper was to summarize the work done on the genus Hemerocallis using genetic related probes in the process of determining the species and their interrelationships, namely using genes to study Hemerocallis systematics. We can reach several conclusions:

1. Use of gene related probes to assess the species in Hemerocallis has commenced. The use of AFLPs seems to be the most regarding at this stage.

2. There exists a multiple set of gene probes which permits the analysis of the genus in an exhaustive manner. Although RFLP and microsatellites and RSS are useful, the AFLP approach allows for massive screening. However mapping the genome is the ultimate goal and then using microarray technology will ensure relationships can be studied in detail.

3. The use of a maximum likelihood approach provides most likely the best tool for assessing genetic heritage and in obtaining trees. This has its weaknesses but still is logically compelling and is as close to what we see in natural processes as well.

4. The use of microarrays and their derivatives will provide the best path to understand mechanisms of gene action between and amongst species.

5. Intra-species and intra-species variations are yet to be determined. Some of the studies focused upon show significant intra-species variation. This must be done in a more exhaustive manner to have better meaning.

7 REFERENCES

1. Atkins, P. **Physical Chemistry**, Freeman (New York) 1990.
2. Baici, A., *Enzyme Kinetics*, The Velocity of Reactions, Biochem Journal, 2006, pp. 1-3.
3. Bartel, B., S. Matsuda, *Seeing Red*, Science, Vol 299, 17 Jan 2003, pp 352-353.
4. Benson, D. L., et al, *Diffusion Driven Instability in an Inhomogeneous Medium*, Bull Math Bio, Vol 55 1993, PP. 365-384.
5. Benson D. L., *Unraveling the Turing Bifurcation Using Spatially Varying Diffusion Coefficients*, Jour Math Bio, Vol 37 1998, pp. 381-417.
6. Berns, R. S., **Principles of Color Technology**, Wiley (New York) 2000.
7. Born, M., E. Wolf, **Principles of Optics**, 4th Ed, Pergamon (New York) 1970.
8. Campbell, A., L. Heyer, **Genomics, Proteomics, and Bioinformatics**, Benjamin Cummings (New York) 2003.
9. Carey, F. A., **Organic Chemistry**, McGraw Hill (New York) 1996.
10. Causton, H. et al, **Microarray Gene Expression and Analysis**, Blackwell (Malden, MA) 2003.
11. Cavalli-Sfroza, L. L., W. F. Bodmer, The Genetics of Human Populations, Dover (Mineola, NY) 1999.
12. Chase, M., Monocot Relationships, an Overview, Journal of Botany, Vol 91 2004 pp 1645-1655.
13. Chen, T., et al, *Modeling Gene Expression with Differential Equations*, Pacific Symposium on Biocomputing, 1999 pp. 29-40.
14. Chung, M., J. Noguchi, *Geographic spatial correlation of morphological characters of Hemerocallis middendorfii complex*, Ann Bot Fennici Vol 35, 1998, pp. 183-189.
15. Chung, M., *Spatial Structure of three Populations of Hemerocallis hakuunensis*, Bot. Bull. Acad. Sci., 2000, Vol 41, pp. 231-236.
16. Cilla, M. L., D. Jackson, *Plasmodesmata Form and Function*, Current Opinion in Cell Bio, Vol 16 2004 pp. 500-506
17. Dahlgren, R.M.T., T**he Families of Monocotyledons**, Springer (New York) 1985.
18. Daly, D., et al, Plant Systematics in the Age of Genomics, Plant Physiology, Dec 2001, pp 1328-1333.
19. Dey, P. M., J. B. Harborne, **Plant Biochemistry**, Academic Press (New York) 1997.
20. Dunn, G., B. Everitt, Mathematical Taxonomy, Dover (Mineola, NY) 2004.
21. Durbin, R. et al, Biological Sequence Analysis, Cambridge (Cambridge) 1998.
22. Durbin M. L. et al, Genes *That Determine Flower Color, Molecular Phylogenetics and Evolution*, 2003 pp. 507-518.
23. Durrett, H., **Color**, Academic Press (New York) 1987.
24. Eisen, M., et al, Cluster Analysis and Display of Genome Wide Expression Patterns, Proc Nat Acad Sci 1998 Vol 98 pp 14863-14868.
25. Erhardt, W., **Hemerocalis**, Timber Press (Portland, OR) 1992.
26. Esau, K., **Anatomy of Seed Plants**, Wiley (New York) 1977.
27. Felsenstein, J., Inferring Phylogenies, Sinauer (Sunderland, MA) 2004.
28. Fox, M. A., J. K. Whitsell, **Organic Chemistry**, Jones and Bartlett (Boston) 1997.
29. Gitzendanner, M., Soltis, P, Patterns of Genetic Variation in Rare and Widespread Plant Congeners, Journal of Botany, Vol 87 2000 pp 783-393.
30. Goodwin, T.W., **Chemistry and Biochemistry of Plant Pigments**, Vols 1 and 2, Academic Press (New York) 1976.
31. Griffiths, A., et al, **Genetic Analysis 5th Ed**, Freeman (New York) 1993.
32. Gusfield, D., Algorithms on Strings, Trees, and Sequences, Cambridge (New York) 1997.
33. Harborne, *Spectral Methods of Characterizing Anthocyanins*, Biochemical Journal, pp 22-28, 1958. http://www.biochemj.org/bj/070/0022/0700022.pdf

34. Harborne, J. B., C. A. Williams, *Anthocyanins and Flavonoids*, Nat Prod Rep 2001 Vol 18 pp. 310-333. http://www.rsc.org/ej/NP/1998/a815631y.pdf
35. Hatzimanikatis, V., *Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information*, Metabolic Engr, Vol 1, 1999, pp. 275-281.
36. Haywood, V. et al, *Plasmodesmata: Pathways for Protein and Ribonucleoprotein Signaling*, The Plant Cell, 2002 PP 303-325.
37. Hildebrand, F. B., **Numerical Analysis**, 2nd Edition, Dover (New York) 1987.
38. Holton, T., E. Cornish, *Genetics and Biochemistry of Anthocyanin Biosynthesis,* The Plant Cell, Vol 7, 1995, pp 1071-1083.
39. Innan, H., et al, A Method for Estimating Nucleotide Diversity from AFLP Data, Genetics, Vol 151 March 1999, pp. 1157-1164.
40. Jaakola, L. et al, *Expression of Genes Involved in Anthocyanin Biosynthesis*, Plant Physiology, Vol 130 Oct 2002, pp 729-739.
41. Jenkins, F. A., H. E. White, **Fundamentals of Optics**, McGraw Hill (New York) 1957.
42. Judd, D. B. et al, **Color** 2^{ND} Edition, Wiley (New York) 1963.
43. Judd, W., et al, Plant Systematics, 3rd Ed, Sinauer (Sunderland, MA) 2008.
44. Judd, W., R. Olmstead, A Survey of Tricolpate (Eudicot) Phylogenetic Relationships, Journal of Botany Oct 2004, pp 1627-1644.
45. Kadar, S., et al, *Modeling of Transient Turing Type Patterns in the Closed Chlorine Dioxide-Iodine-Malonic Acid-Starch Reaction System*, J Phys Chem, Vol 99 1995 pp. 4054-4058.
46. Kohane, I., et al, **Microarrays for an Integrative Genomics**, MIT Press (Cambridge) 2003.
47. Koopman, W., et al, AFLP Markers as a Tool to Reconstruct Complex Relationships: A Case Study in Rosa, Journal of Botany, Vol 95 2008 pp 353-366.
48. Lee, D., **Nature's Palette**, University of Chicago Press (Chicago) 2007.
49. Lesk, A., **Bioinformatics**, Oxford (New York) 2002.
50. Levi, L., **Applied Optics**, Wiley (New York) 1968.
51. Mauseth, J. D., **Plant Anatomy**, Benjamin (Menlo Park, CA) 1988.
52. Mayr, E., Toward a New Philosophy of Biology, Harvard University Press (Cambridge) 1988.
53. Mayr, E., Evolution and the Diversity of Life, Harvard University Press (Cambridge) 1976.
54. Mayr, E., Population, Species and Evolution, Harvard University Press (Cambridge) 1970.
55. Mayr, E., The Growth of Biological Thought, Harvard University Press (Cambridge) 1982.
56. 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
57. McGarty, T., **Stochastic Systems and State Estimation**, Wiley (New York) 1974.
58. 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 .
59. 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
60. McMurry, J., Begley, T., The **Organic Chemistry of Biological Pathways**, Roberts & Company Publishers, 2005.
61. Milgrom, L. R**., The Colours of Life**, Oxford (New York) 1997.
62. Mohr, H., P. Schopfer, **Plant Physiology**, Springer (New York) 1995.
63. Mol, J, et al, *How Genes Paint Flowers and Seeds, Trends in Plant Science*, Vol 3 June 1998, pp 212-217.
64. Mol, J., et al, *Novel Colored Plants*, Current Opinion in Biotechnology, Vol 10, 1999, pp 198-201.
65. Mueller, U., L. Wolfenbarger, AFLP Genotyping and Fingerprinting, Trends Ecol. Evol. **14: 1999, pp** 389–394. .
66. Munson, R., **Hemerocalis, The Daylily**, Timber Press (Portland, OR) 1989.
67. Murray, J., **Mathematical Biology**, Springer (New York) 1989.
68. Murrell, J., *Understanding Rate of Chemical Reactions*, University of Sussex.
69. Naik, P. S., et al, *Genetic manipulation of carotenoid pathway in higher plants*, Current Science, Vol 85, No 10, Nov 2003, pp 1423-1430.
70. Nei, M., S.. Kumar, Molecular Evolution and Phylogenetics, Oxford (New York) 2000.
71. Nicklas, K. J., The Bio Logic and Machinery of Plant Morphogenesis, Am Jour Bot 90(4) 2003 pp. 515-525.

72. Noguchi, J., H. De-yuan, *Multiple origins of the Japanese nocturnal Hemerocalis citrina*, Int Jrl Plant Science, 2004, Vol 16, pp. 219-230.
73. Norton, J., *Some Basic Hemerocallis Genetics*, American Hemerocallis Society, 1982.
74. Norton, J. 1972. *Hemerocallis* Journal 26 (3) in Bisset, K. 1976. *Spectrophotometry, Chromatography and Genetics of Hemerocallis Pigments*. Dissertation, Florida State Univ.
75. Oparka, K. J., A. G. Roberts. Plasmodesmata*, A Not So Open and Shut Case*, Plant Phys, Jan 2001, Vol 125 pp. 123-126.
76. Percus, J., Mathematics of Genome Analysis, Cambridge (New York) 2002.
77. Perkins, T., et al, *Inferring Models of Gene Expression Dynamics*, Journal of Theoretical Biology, Vol 230, 2004, pp. 289-299.
78. Petit, T. *The Patterned Daylily*, The Daylily Journal, Vol 62 No 2 2007 pp. 125-141.
79. Planet, P. et al, Systematic Analysis of DNA Microarray Data: Ordering and Interpreting Patterns of Gene Expression, Cold Spring Harbor, Genome Research, 2001 pp 1149-1155.
80. Rossi, B., **Optics**, Addison Wesley (Reading, MA) 1957.
81. Saitou, N., M. Nei, The Neighbor Joining Method: A New Method for Reconstructing Phylogenetic Trees, Molecular Biological Evolution, Vol 4 1987 pp 406-425.
82. Sansone, G et al, **Orthogonal Functions**, Academic Press (Mew York) 1958.
83. Sattath, S., A. Tversky, Additive Similarity Trees, Psychometrica, Vol 42, 1977 pp. 319-345.
84. Schnell. S, T. Turner, *Reaction Kinetics in Intracellular Environments with Macromolecular Crowding*, Biophys and Molec Bio vol 85 2004 pp. 235-260.
85. Sears, F. W., **Optics**, Addison Wesley (Reading, MA) 1949.
86. Sokal, R., P. Sneath, Principle of Numerical Taxonomy, Freeman (San Francisco) 1963.
87. Soltis, D. et al, Evolution of Genome Size in Angiosperms, Journal of Botany, Nov 2003, pp 1596-1603
88. Somasundaram, S., M. Kalaiselvam, Molecular Tools for Assessing Genetic Diversity, UN University Course, http://ocw.unu.edu/international-network-on-water-environment-and-health/unu-inweh-course-1-mangroves/Molecular_Tools__for_Assessing_Genetic_Diversity.pdf .
89. Stace, C., Plant Taxonomy and Biosystematics, Arnold (London) 1989.
90. Stout, A.B., **Daylilies**, Saga Press (Millwood, NY) 1986.
91. Strong, J., **Concepts of Classical Optics**, Freeman (San Francisco) 1958.
92. Studier, J., K. Kappler, A Note on the Neighbor Joining Algorithm, Molecular Biological Evolution, Vol 5 1988 pp 729-731.
93. Szallasi, Z. **System Modeling in Cellular Biology: From Concepts to Nuts and Bolts**. MIT Press (Cambridge) 2006.
94. Taiz, L., E. Zeiger, **Plant Physiology**, Benjamin Cummings (Redwood City, CA) 1991.
95. Taubes, C. H., Modeling Differential Equations in Biology, Cambridge (New York) 2001.
96. Tetter, A., et al, Primer on Medical Genomics Part III, Mayo Clinic Proc, Vol 77 2002 pp 927-940.
97. Tinoco, I. et al, **Physical Chemistry**, Prentice Hall (Englewood Cliffs, NJ) 1995.
98. Tobias, A., *Directed Evolution of Biosynthetic Pathways to Carotenoids with Unnatural Carbon Bonds*, PhD Thesis, Cal Tech, 2006.
99. Tomkins, J. R., *DNA Fingerprinting in Daylilies*, Parts I and II, Daylily Journal, Vol 56 No 2 and 3 2001, pp. 195-200 and pp. 343-347.
100. Tomkins, J. R., *How much DNA is in a Daylily*, Daylily Journal, Vol 58 No 2 2003, pp. 205-209.
101. Tomkins, J., et al, *Evaluation of genetic variation in the daylily (Hemerocalis) using AFLP markers*, Theor Appl Genet Vol 102, 2001, pp. 489-496.
102. Turing, A., *The Chemical Basis of Morphogenesis*, Phil Trans Royal Soc London B337 pp 37-72, 19459.
103. Van Trees, H. L., **Detection, Estimation and Modulation Theory**, Wiley (New York) 1968.
104. Vohradsky, J., *Neural Network Model of Gene Expression*, FASEB Journal, Vol 15, March 2001, pp. 846-854.
105. Vos, P., et al, AFLP: A New Technique for DNA Fingerprinting, Nuclear Acids Research, Vol 23 1995 pp 4407-4414.
106. Wade, L. G., **Organic Chemistry**, Prentice Hall (Saddle River, NJ) 2003.
107. Watson, J., et al, **Molecular Biology of the Gene**, Benjamin Cummings (San Francisco) 2004.
108. Watson, J. et al, Recombinant DNA, 3rd Ed, Freeman (New York) 2007.
109. Wen, X., et al, Large Scale Temporal Gene Expression Mapping of Central Nervous System Development, Proc Nat Acad Sci Neurology 1998 Vol 95 pp 334-339.
110. Winkel-Shirley, B., *Flavonoid Biosynthesis*, Plant Physiology, Vol 126 June 2001 pp 485-493. http://www.plantphysiol.org/cgi/reprint/126/2/485

111. Yu, O. et al, *Flavonoid Compounds in Flowers: Genetics and Biochemistry*, General Science Books, http://www.danforthcenter.org/yu/pdf/e-flower-2006.pdf

112. Zambryski, P., C*ell to Cell Transport of Proteins and Fluorescent Tracers via Plasmodesmata During Plant Development*, Jour Cell Bio Vol 161 No 2 Jan 2004, pp. 165-168.