Training program on

Bioinformatics Tools for Genomics Research

February 14-18, 2011

Coordinator
Dr. R. Samiyappan
Director (CPMB)

Co-coordinator
Dr. J. Ramalingam
Associate Professor

Compiled by
Dr. V. Udayasuriyan, Professor & Head (DPMB&B)
N. Bharathi, Assistant Professor

Distributed Information Sub Centre (DISC)
Department of Plant Molecular Biology and Biotechnology
Centre for Plant Molecular Biology
Tamil Nadu Agricultural University
Coimbatore- 641003

TABLE OF CONTENTS

Applications of Bioinformatics in Plant Virus Research

Dr. R. Usha, Professor,
Department of Plant Biotechnology, School of Biotechnology,
Madurai Kamaraj University, Madurai -625021.
abayamba@gmail.com

Plant viruses cause enormous losses to crops worldwide. In order to devise strategies for disease control, it is essential to understand the sequences of the plant viral genomes and the proteins encoded by them. Currently there are several thousand plant viral sequences in the databases. Bioinformatics plays an important role in mining valuable information from these data. The tools of bioinformatics are used right from the beginning, namely the design of PCR primers for the amplification of plant viral genomes. Determining the phylogenetic relationships between viruses, elucidating the recombination, molecular reassortment and mutations in plant viruses, choice of viral genes for genetic engineering of plants for developing virus-resistant plants, display of foreign epitopes on the surface of plant virus coat proteins are all processes, which rely on bioinformatics.

Examples for the applications of bioinformatics in plant virus research from our lab:

- Classification of *Cardamom mosaic virus* (CdMV) as a new member of the *Macluravirus* genus of the *Potyviridae* family based on the sequence analysis of the coat protein.
- Phylogenetic relationships between the various isolates of CdMV gave key information about the movement of the virus along the cardamom tract.
- From the analysis of the sequence of the surface exposed regions of CdMV, experiments for the insertion of epitopes from HIV could be devised.
- Identification of a novel satellite DNA associated with Bhendi yellow vein mosaic disease.

- Detection of recombination and molecular reassortment among the viruses causing yellow mosaic disease in legumes in South and South East Asia using multiple sequence alignment and recombination detection programs.
- Detection of motifs and domains in the replication associated protein and movement protein of the soybean virus and the coat protein, C2, C4 and βC1 proteins of the bhendi virus has led to the elucidation of the functions of these proteins.
- Promoter regions in *Bhendi yellow vein mosaic virus* (BYVMV) could be analyzed.
- Host pathogen interactions involving BYVMV could be studied with experiments designed using bioinformatics tools.
- Identification of *Vernonia yellow vein virus* and *Horsegram yellow mosaic virus* as two new species of *Begomovirus* genus of *Geminiviridae* family.
- The sequences of candidate genes from the cardamom, bhendi and soybean viruses could be chosen for genetic engineering of plants for virus resistance.
- Virus induced gene silencing vectors could be devised for functional genomics of plants.
- Molecular modeling of BYVMV coat protein could be carried out with the help of bioinformatics tools.

An Overview of Prokaryotic & Eukaryotic Genomes

Dr. V. Udayasuriyan, Professor & Head,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
biotech@tnau.ac.in

How genomic organization varies in different organisms – from viruses to bacteria to eukaryotes will undoubtedly provide a better understanding of the evolution of organisms on Earth. We will first survey what is known about chromosomes in viruses and bacteria. Then, we will examine some of the basic questions of eukaryotic genomic organization.

Prokaryotic chromosomes

The chromosomes of viruses and bacteria are much less complicated. They usually consist of single nucleic acid molecule, largely devoid of associated proteins and containing relatively little genetic information in comparison to the multiple chromosomes constituting the genome of higher forms. The chromosomes of viruses consist of a nucleic acid molecule – either DNA or RNA – that can be either single or double stranded. They can exist as circular structures (covalently closed circles), or they can take the form of linear molecules. Bacterial chromosomes are also relatively simple in form. They always consist of a double – stranded DNA molecule, compacted into a structure sometimes referred to as the nucleoid. *Escherichia coli*, the most extensively studied bacterium, has a large, circular chromosome measuring approximately 1.2 mm in length.

Eukaryotic chromosomes

Following chromosome separation and cell division, cells enter the interphase stage of the cell cycle, during which time the components of the chromosome uncoil and are present in the form referred to as chromatin. While in interphase, the chromatin is dispersed in the nucleus, and the DNA of each chromosome is replicated. As the cell cycle progresses, most cells reenter mitosis, where upon chromatin coils into visile

chromosomes once again. This condensation represents a contraction in length of some 10,000 times for each chromatin fiber. The DNA in the *E. coli* chromosome is 1.2 mm long, the DNA in each human chromosome ranges from 19 to 73 mm in length. In a single human nucleus, all 46 chromosomes contain sufficient DNA to extend almost 2 meters. This genetic material, along with its associated proteins, is contained within a nucleus that usually measures about 5-10 μm in diameter.

Chromatin structure and nucleosomes

The genetic material of viruses and bacteria consists of strands of DNA or RNA nearly devoid of proteins. In eukaryotic chromatin, a substantial amount of protein is associated with the chromosomal DNA in all phases of the eukaryotic cell cycle. The associated proteins are divided into basic, positively charged histones and less positively charged nonhistones. Of the proteins associated with DNA, the histones clearly play the most essential structural role. Electron microscopic observations of chromatin have revealed that chromatin fibers are composed of linear arrays of spherical particles. Discovered by Ada and Donald Olins, the particles occur regularly along the axis of a chromatin strand and resemble beads, are now called nucleosomes.

Repetitive DNA

In addition to single and multiple copies of unique DNA sequences that make up genes, a great deal of the DNA sequences within chromosomes is repetitive in nature and most repetitive sequences serve no known function. In humans, it appears that the estimated 30,000 functional genes occupy less than 5 per cent of the genome. So far 41 eukaryotic genomes and 1442 prokaryotic genomes have been completed. The size of representative genome is given in Table 1.

Gene expression

The first step in gene expression involves the transfer of information present on one of the two strands of DNA (the template strand) into an RNA complement through the process of transcription. Once synthesized, this RNA acts as a “messenger” molecule, bearing the coded information – hence its name, messenger RNA (mRNA). Such RNAs

then associate with ribosomes, in which decoding into proteins occurs. The genetic code is written in units of three letters – ribonucleotides present in mRNA that reflect the stored information in genes. Each triplet code word directs the incorporation of a specific amino acid into a protein as it is synthesized. The code is nearly universal. With only minor exceptions, a single coding dictionary is used by almost all viruses, prokaryotes, archaea and eukaryotes. AUG encodes methionine, which intiates most polypeptide chains. All other amino acids except tryptophan, which is encoded only by UGG, are represented by two to six triplets. The triplets UAA, UAG and UGA are termination signals and do not encode any amino acids. Each triplet codon in the mRNA is, in turn, complementary to the anticodon region of its corresponding tRNA as the amino acid is correctly inserted into the polypeptide chain during translation.

To prove that RNA can be synthesized on a DNA template, it was necessary to demonstrate that there is an enzyme capable of directing this synthesis. By 1959, several investigators, including Samuel Weiss had independently discovered such a molecule from rat liver. Called RNA polymerase, it has the same general substrate requirements as does DNA polymerase, the major exception being that the substrate nucleotides contain the ribose rather than the deoxyribose form of the sugar.

The initial step in transcription is referred to as template binding. In bacteria, the site of this initial binding is established when the σ subunit of RNA polymerase recognizes specific DNA sequences called promoters. These regions are located in the 5' region, upstream from the point of initial transcription of a gene. Because the interaction of promoters with RNA polymerase governs transcription, the nature of the binding between them is at the heart of discussions concerning genetic regulation. The first is the concept of consensus sequences of DNA. Two such sequences have been found in bacterial promoters. One, TATAAT, is located 10 nucleotides upstream from the site of initial transcription (the – 10 region, or Pribnow box). The other TTGACA, is located 35 nucleotides upstream (the – 35 region). Mutations in both regions diminish transcription, often severely. In most eukaryotic genes studied, a consensus sequence comparable to that in the –10 region has been recognized. Because it is rich in adenine and thymine residues, it is called the TATA box. The second point is that the degree of RNA

polymerase binding to different promoters varies greatly, which causes the variable gene expression mentioned earlier. Currently, this is attributed to sequence variation in the promoters.

It is important to note that bacteria groups of genes whose protein products are involved in the same metabolic pathway are often clustered together along the chromosomes. In many such cases, the genes are contiguous and all but the last gene lack the encoded signals for termination of transcription. The result is that during transcription a large mRNA is produced, encoding more than one protein. Since genes in phage and bacteria were historically referred to as cistrons, the RNA is called a polycistronic mRNA. Since the products of genes transcribed in this fashion are usually all needed at the same time, this is an efficient way to transcribe and, subsequently, to translate the needed genetic information. In eukaryotes, monocistronic mRNAs are the rule.

Transcription in eukaryotes differs from prokaryotes

1. Transcription in eukaryotes occurs within the nucleus under the direction of three separate forms in RNA polymerase. In eukaryotes, the RNA transcript is not free to associate with ribosomes prior to the completion of transcription. For the mRNA to be translated, it must move out of the nucleus into the cytoplasm.
2. In addition to promoters, other control units, called enhancers, may be located in the 5' –regulatory region upstream from the initiation point, but they have also been found within the gene or even in the 3' downstream region beyond the coding sequence.

3. Maturation of eukaryotic mRNA from the primary RNA transcript involves many complex stages referred to generally as "processing". An initial processing step involves the addition of a 5' cap and a 3' tail to most transcripts destined to become mRNAs. The initial (or primary) transcripts are most often much larger than those that are eventually translated. Sometimes called pre-mRNAs, they are part of a group of molecules found only in the nucleus – a group referred to collectively as heterogeneous nuclear RNA (hnRNA). In those that are converted, substantial amounts of the ribonucleotide sequence are excised and the remaining segments are

spliced back together prior to translation. This phenomenon has given rise to the concepts of split genes and splicing in eukaryotes.

Coding regions of eukaryotic genes are interrupted

The internal DNA sequences are present in initial RNA transcripts, but they are removed before the mature mRNA is translated. Such nucleotide segments are called intervening sequences, and the genes that contain them are known as split genes. Those DNA sequences that are not represented in the final mRNA products are also called introns ("int" for intervening), and those retained and expressed are called exons ("ex" for expressed). Splicing involves the removal of the ribonucleotide sequences present in introns as a result of an excision process and the rejoining of exons. The *pro-α-2(1) collagen* gene contains 50 introns. Only about 15 per cent of the collagen gene consists of exons that finally appear in mRNA.

Organelle genome

Most of the organelle genomes characterized in eukaryotes are in the form of a single circular molecule (denoted mtDNA in the mitochondrion and ctDNA in the chloroplast). Usually there are several copies of the genome in the individual organelle. Since there are multiple organelles per cell, there is a large number of organelle genomes per cell. Chloroplast genomes are relatively large, about 140 kb in higher plants. Animal cells have small mitochondrial genomes of about 16.5 kb. Plants show wide range of variation in mitochondrial DNA size, with a minimum of ~100kb.

Table 1. Size of representative genomes

S.No	Species	Type of Organism	Genome size (Mb)
1.	Mycoplasama genitalium	Bacterium	0.58
2.	Haemophilus influenza	Bacterium	1.83
3.	Escherichia coli	Bacterium	4.64
4.	Saccharomyces cerevisiae	Yeast	12.10
5.	Caenorhabditis elegans	Nematode worm	97.00
6.	Drosophila melanogaster	Insect	180.00
7.	Arabidopsis thaliana	Plant	125.00
8.	Homo sapiens	Mammal	3200.00
9.	Triticum aestivum	Plant (Wheat)	17000.00
10.	Oryza sativa	Plants	334.76

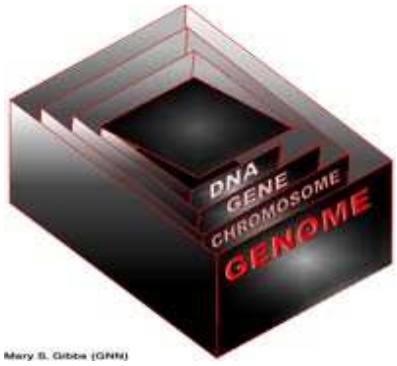
Plant Genomics - An overview

Dr. P. Nagarajan, Professor,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
p_nrajan@yahoo.com

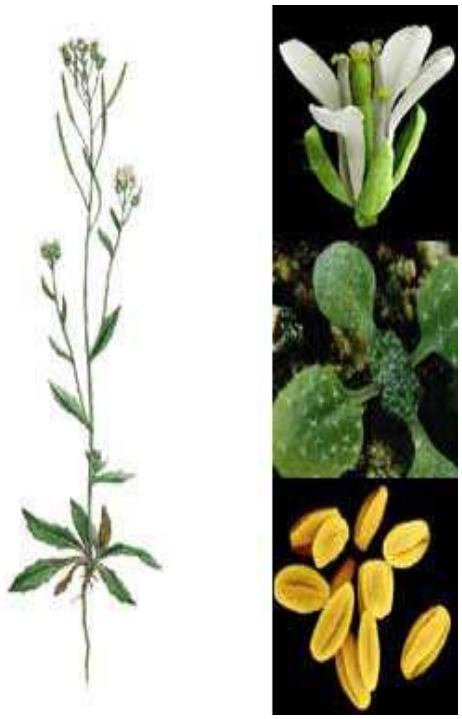
Introduction

In modern molecular biology and genetics, the genome is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA. The term was adapted in 1920 by Hans Winkler, Professor of Botany at the University of Hamburg, Germany. It is a portmanteau word *i.e.,* blending the sounds and the meaning of the two other words ***gene*** and *chromos**ome,*** (Gene + ome (for mass), large number of genes gathered. The term genome can be applied specifically to mean that stored on a complete set of *nuclear DNA* (i.e., the "nuclear genome"). Although plants possess mitochondrial and chloroplast genomes, their nuclear genome is the largest and most complex. Until very recently, the molecular analysis of plants often focused on the single gene level. Recent technological advances have changed this paradigm, enabling the analysis of organisms in terms of genome organization, expression and interaction. The study of the way genes and genetic information are organized within the genome, the methods of collecting and analyzing this information, and how this organization determines their biological functionality is referred to as genomics. Plant genomics is reversing the previous paradigm of identifying genes behind biological functions and instead focuses on finding biological functions behind genes and also reduces the gap between phenotype and genotype. In this review genomic issues are addressed from a plant perspective and is organized into two main sections. The first deals with the current understanding of plant genomes, their genetic structure and how whole genomes are sequenced, and its second section addresses some approaches used in order to achieve the final aim of genomics: finding the biological and functional significance of DNA sequence.

Plant genome organization



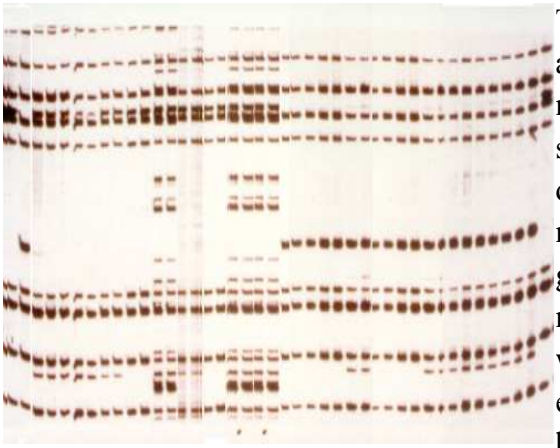
Plant genomes are best described in terms of genome size, gene content, extent of repetitive sequences and polyploidy/duplication events. There is extensive variation in nuclear genome size without obvious functional significance of such variation. Plant genomes contain various repetitive sequences, long interspersed nuclear elements and short-interspersed nuclear elements. It is widely accepted that 70-80% of flowering plants are the product of at least one polyploidization event (Barnes, 2002). Many economically important plant species, such as corn, wheat, potato, and oat are either ancient or more recent polyploids, comprising more than one, and in cases such as wheat, three different homologous genomes within a single species. Duplicated segments also account for a significant fraction of the rice genome. About 60% of the *Arabidopsis* genome is present in 24 duplicated segments, each more than 100 kilobases (kb) in size. Model organisms (*Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae*) provide genetic and molecular insights into the biology of more complex species. Since the genomes of most plant species are either too large or too complex to be fully analyzed, the plant scientific community has adopted model organisms. Although the advent of tissue culture techniques fostered the use of tobacco and petunia, the species now used as model organisms for mono-and dicotyledonous plants are rice (*Oryza sativa*) and *Arabidopsis* (*Arabidopsis thaliana*) respectively.



Arabidopsis, a small Cruciferae plant without agricultural use, sets seed in only 6 weeks from planting, has a small genome of 120 Megabases (Mb) and only five chromosomes. There are extensive tools available for its genomic analysis, whole genome sequence, Expressed Sequence Tags (ESTs) collections, characterized mutants and large populations mutagenized with insertion elements (transposons or the T-DNA of *Agrobacterium*). *Arabidopsis* can be genetically transformed on a large scale with *Agrobacterium tumefaciens* and biolistics. Other tools available for this model plant are saturated genetic and physical maps.

Unlike *Arabidopsis*, rice is one of the world's most important cereals. More than 500 million tons of rice is produced each year, and it is the staple food for more than half of the worldís population. There are two main rice subspecies. *Japonica* is mostly grown in Japan, while *indica* is grown in China and other Asia-Pacific regions. Rice also has very saturated genetic maps, physical maps, whole genome sequences, as well as EST collections pooled from different tissues and developmental stages. It has 12 chromosomes, a genome size of 420 Mb, and like *Arabidopsis*, it can be transformed through biolistics and *A. tumefaciens*. Efficient transposon-tagging systems for gene knockouts and gene detection have not yet become available for saturation mutagenesis in rice, although some recent successes have been reported.

Molecular markers



The development of molecular markers has allowed for constructing complete genetic maps for most economically important plant species. They detect genetic variation directly at the DNA level. A myriad of molecular marker systems are available. A genetic map represents the ordering of molecular markers along chromosomes as well as the genetic distances, generally expressed as centiMorgans (cM), existing between adjacent molecular markers.

Genetic maps in plants have been created from many experimental populations, but the most frequently used are F2, backcrosses and recombinant inbred lines. Although longer to develop, recombinant inbred lines offer a higher genetic resolution and practical advantages. Genetic maps contribute to the understanding of how plant genomes are organized and once available they facilitate the development of practical applications in plant breeding, such as the identification of Quantitative Trait Loci and Marker Assisted Selection. Quantitative Trait Loci analysis refers to the identification of genomic regions associated with the phenotypic expression of a given trait. Once the location of such genomic regions is known they can be assembled into designer genotypes, i.e. individuals carrying chromosomic fragments associated with the expression of a given phenotype. Genetic maps are also an important resource for plant gene isolation, as once the genetic position of any mutation is established, it is possible to attempt its isolation through positional cloning (Campos-de Quiroz et al., 2000). Furthermore, genetic maps help establish the extent of genome colinearity and duplication between different species.

Physical maps

Although genetic maps provide much-needed landmarks along chromosomes, they are still too far apart to provide an entry point into genes, since even in model plants the kilobases per centiMorgan (kb/cM) ratio is large, from 120 to 250 kb/cM in *Arabidopsis* and between 500 and 1.500 kb/cM in corn. Therefore, a 1 cM interval may harbor ~30 to 100 or even more genes. Physical maps bridge such gaps, representing the

entire DNA fragment spanning the genetic location of adjacent molecular markers. Physical maps can be defined as a set of large insert clones with minimum overlap encompassing a given chromosome. First generation physical maps in plants were based on YACs (Yeast Artificial Chromosomes). Then with Bacterial Artificial Chromosomes (BACs) and P1-derived artificial chromosomes. Although BAC vectors are relatively small (molecular weight of BAC vector pBeloBAC11 is 7.4 kb for instance), they carry inserts between 80 and 200 kb on average and possess traditional plasmid selection features such as an antibiotic resistance gene and a polycloning site within a reporter gene allowing insertional inactivation. Physical and genetic maps can be aligned, bringing along continuity from phenotype to genotype. Physical maps provide the bridge needed between the resolution achieved by genetic maps and that needed to isolate genes through positional cloning.

Genome co linearity/Genome evolution

A remarkable feature of plant genomics is its ability to bring together more than one species for analysis. The comparative genome mapping of related plant species has shown that the organization of genes is highly conserved during the evolution of members of taxonomic families. This has led to the identification of genome colinearity between the well-sequenced model crops and their related species (e.g. *Arabidopsis* for dicots and rice for monocots). Colinearity overrides the differences in chromosome number and genome size and can be defined as conservation of gene order within a chromosomal segment between different species. A related concept is synteny, which refers to the presence of two or more loci on the same chromosome regardless they are genetically linked or not. Colinear relationships have been observed among cereal species (corn, wheat, rice, barley), legumes (beans, peas and soybeans), pines and *Cruciferae* species (canola, broccoli, cabbage, *Arabidopsis thaliana*). Comparing sequences of soybean and *Arabidopsis* demonstrated partial homology between two soybean chromosomes and a 25 cM section of chromosome 2 from *Arabidopsis*. Colinearity has also been established between rice and most cereal species, allowing the use of rice for genetic analysis and gene discovery in genetically more complex species, such as wheat

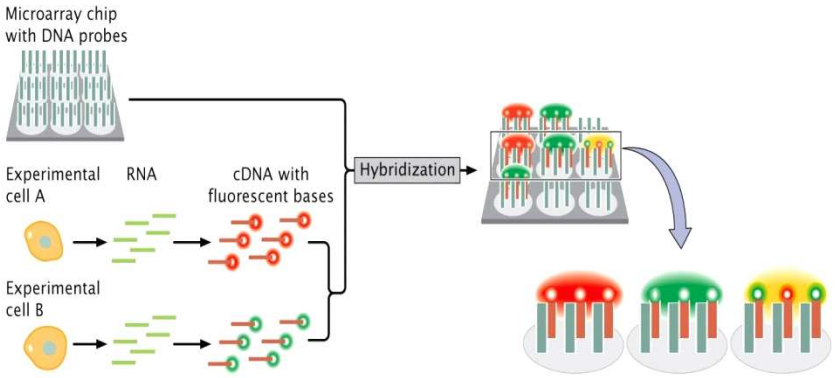
and barley. General gene structure was largely conserved between rice and barley. There are even reports of colinearity across the mono-dicotyledoneous division involving *Arabidopsis* and cereals which diverged as far back as 200 million years ago. Exploiting colinearity helps to establish cross-species genetic links and also aids in the extrapolation of information from species with simpler genomes (i.e. rice) to genetically complex species (corn, wheat). Furthermore, it reflects the power of genomics to integrate genetic information across species.

Whole genome sequencing

Genetic and physical maps at the inter- or intra-species level represent a key layer of genomic information. However, sequence data represents the ultimate level of genetic information. There are two main approaches to large scale sequencing. In clone-by-clone strategies and shotgun approaches. The *Arabidopsis* genome was the first to be fully sequenced, approximately one third of the genes putatively identified in *Arabidopsis* encode products lacking significant similarity to proteins of known function in other organisms. Moreover, only 9% of its genes have been characterized experimentally. In rice, the IRGSP (International Rice Genome Sequencing Project) started in 1997, and includes members from developing countries in addition to European and USA partners. Syngenta and a Chinese group recently made available the sequences of japonica and *indica* rice, respectively. Nevertheless, the actual number of genes existing in *Arabidopsis*, rice, or any other sequenced species remains to be established through functional genomic experiments that establish the biological meaning of DNA sequences, since gene prediction through homology comparisons and software tools is a statistical "best informed guess" rather than a biologically based process. Availability of extensive EST collections, which now exist in several plant species, including corn and soybean, reduce the dependence of the annotation process on computational gene predictions. The field of genomics that addresses the function of genes discovered through sequencing efforts is referred to as *functional genomics*. Since genes encoding traits are expected to be functional across species, most of the information thus gathered will be useful to address plant improvement issues or biological processes.

ESTs (Expressed Sequence Tags)

Large scale sequencing facilities allow the development of ESTs, also known as single pass sequences of random cDNAs. Since many genes have tissue-specific temporal expression patterns, in order to collect cDNAs from most expressed genes it is necessary to prepare cDNA libraries from several different tissues and also from tissues challenged with diverse biotic and abiotic factors. Searches for homology with known genes from other species allows the assignment of putative biological functions to ESTs. The use of ESTs also permits the identification of genes encoding functionally unknown proteins. There is an extensive collection of public ESTs from a number of organisms in the database dbEST, a division of GenBank. Furthermore, they have provided the platform necessary for transcriptional profiling experiments. EST sequencing programs, therefore, provide a powerful lead into genomic approaches in plants. This area of genomics involves the study of gene expression patterns across a wide array of cellular responses, phenotypes and conditions. There are several systems available to analyze the parallel expression of many genes such as macroarrays and Serial Analysis of Gene Expression (SAGE)), which consists of identifying short sequence tags from individual transcripts, their concatenation, sequencing and subsequent digital quantitation. SAGE provides expression levels for many transcripts across different stages of development.



There are open and closed transcriptional profiling systems. Open technologies, One example of such a system is the GeneCalling technology). Another open system is provided by Massively Parallel Sequence Signatures (MPSS), where microbeads are used to construct libraries of DNA templates and create hundred of thousands of gene

signatures). Closed systems, on the other hand, analyze genes that have been previously characterized. They include most of the diverse microarray systems available, Microarray applications are broadly classified as expression-specific and genome-wide expression studies. The value of using microarrays to identify novel response genes has been demonstrated by studying the gene expression patterns during corn embryo development (Lee et al., 2002), the response to drought and cold stresses, herbivory, and nitrate treatments. This principle revealed previously unexpected relationships between low soil phosphate levels and cold acclimation in *Arabidopsis.* Transcriptional profiling technologies play a central role in predicting gene function since sequence comparison alone is insufficient to infer function. Unlike animals, plants cannot move and have developed exquisite mechanisms to cope with changing environmental conditions and biotic challenges, since these directly or indirectly affect most biological processes occurring in plants. Therefore, a significant proportion of the information gathered by specific and genome wide transcription profiling processes should have practical applications and facilitate the development of plants more resilient to biotic and abiotic stimuli. There are unsupervised approaches in which no knowledge about how the genes assessed are organized is available, for instance clustering algorithms, principal component analysis and *k*-means clustering. There are also neural network-based methods such as Self Organizing Maps (SOMs). The extensive amount of information generated from microarray experiments is best managed with Laboratory Information Management Systems handling sample submission, sample processing, sample tracking, data retrieving, sorting, visualization and statistical analysis. Observations of expression data may help generate hypotheses, but additional experimentation, for example genetic mapping or transgenesis, may be necessary to validate these hypotheses.

Conclusion

The large number of genes handled simultaneously by genomics sets a new paradigm in plant biology, since it allows the genetic integration of diverse processes, tissues and organisms. It is expected that a significant proportion of such information will be transferred to plant improvement programs and will thus contribute to meeting the

increasing food requirements of the world. Plant genomics will revolutionize the study of the molecular basis of plant biology. The traditional hypothesis-driven approach will be gradually transformed into an unbiased data collection at the tissue/organism level approach followed by bioinformatic analyses. Finally, genomics is the ultimate interdisciplinary approach, as it covers the entire spectrum from DNA sequencing to field-based research. Only through the integrated endeavor of genetics, biology, bioinformatics, molecular biology, engineering, microbiology and related fields will the extensive benefits of genomics to mankind become reality.

References

Barnes, S., 2002. Comparing Arabidopsis to other flowering plants. *Curr Op Plant Biology* 5,128-133.

Campos, D. E., Quiroz, H., Magrath, R., Mccallum, D., Kroymann, J., Scnabelrauch. D., Mitchell-Olds, T., Mithen, R. 2000 Keto acid elongation and glucosinolate biosynthesis in Arabidopsis thaliana. *Theor Appl Genet* 101, 429-437.

Lee, J.M., Williams M.E., Tingey S.V., Rafalski J.A. 2002. DNA array profiling of gene expression changes during maize embryo development. *Funct Integr Genomics* 2, 13-27.

Rafalski, A.J. 2002 Plant genomics: present state and a perspective on future developments. *Briefings in Fundamental Genomics and Proteomics* 1, 1-15.

DNA Markers and Mapping Major and Minor Genes

Dr. M. Maheswaran, Professor,
CPBG, TNAU, Coimbatore 641003.
mahes@tnau.ac.in

Introduction

Advances in molecular biology during the last three decades have provided new classes of genetic markers at the level of DNA. Some of the most routinely used markers include Restriction Fragment Length Polymorphism (RFLP), Randomly Amplified Polymorphic DNA (RAPD), Sequence Tagged Sites (STS), Expressed Sequence Tags (EST), Sequence Characterized Amplified Regions (SCAR), Simple Sequence Repeats (SSR) and Amplified Fragment Length Polymorphism (AFLP). The unlimited availability of all these markers enabled plant geneticists and breeders to establish well saturated genetic maps for several crop species. These developments have stimulated new interest in exploring the applications of genetic markers in plant breeding. One among the applications is mapping genes of major and minor nature in crop plants.

Gene Mapping

Molecular markers offer a tool for locating genes governing agronomically important characters *via* linkage to mapped DNA sequences. Phenotypic evaluation at the whole plant level or at the cellular level provides information which can be used to determine the chromosomal location of the genes that confer the phenotype of interest. This is accomplished by analyzing linkage between mapped molecular markers and expression of the target phenotype in a range of related individuals. Markers linked to the genes of interest function as "genetags" facilitating selection of favourable alleles in a breeding programme.

Like for linkage map construction, gene tagging component also needs a suitable population in which the trait to be tagged with molecular markers shows clear-cut segregation with a higher level of polymorphism for the molecular markers. The process of gene tagging involves two steps: l) surveying parents with molecular markers for their

level of polymorphism, and 2) surveying the polymorphic markers on progenies with an aim to tag the trait of interest with a molecular marker(s).

Mapping major genes.

Establishing associations between molecular markers and simply inherited traits is comparatively easier. To date, several major genes have been tagged with molecular markers and among them genes conferring resistance to pest and diseases are more common. The strategies followed to tag major genes are:

1. For most of the traits, the genetics is not an established phenomenon. As far as crop species are concerned, the genetics of agronomically important traits has not been studied, except for the traits such as resistance to various insects and diseases. Under these circumstances, if one wishes to tag genes of a trait, the general process of parental survey followed by progeny survey can be followed. This is a time consuming process as far as tagging major genes is concerned.

2. When a breeder understands the genetics of a trait, the next step to be followed is the construction of Near-Isogenic lines. Near isogenic lines (NILs) are constructed *via* the back cross breeding method. These NILs possess constant donor parent (DP) DNA introgressed into the recurrent parent (RP). The strategy involves survey of RP, NIL and DP with molecular markers to identify putative positives (markers present in the DP and NIL, but absent in RP). Then the putative positives are surveyed on a segregating population of RP and NIL to establish linkage between a marker and the gene of interest. Thus near Isogenic lines remain as a potential resource in the molecular marker approach (Muehlbauer *et al*., l988, Martin *et al.,* 1991).

3. Construction of NILs is a laborious and time consuming process. An alternative strategy has been proposed by Michelmore et al., (l99l) and named Bulked Segregant Analysis (BSA). In a population segregating for a trait of interest, individuals are sorted into two groups based on the expression of that trait. The two groups will differ at loci linked to the trait, but will be randomly segregating for unlinked regions. DNA samples from each group are pooled separately, and the two DNA pools are screened for markers polymorphism. Markers polymorphic between the two pools are expected to be linked to

the trait of interest. This is a viable strategy to a variety of simply inherited traits to tags their genes with molecular markers.

Mapping genes controlling quantitative traits

In crop breeding, most of the traits that breeders concerned with are polygenically controlled. Location of polygenes in individuals by conventional analysis was difficult. The advent of molecular marker technology provides the geneticists with powerful new tools for identifying the component Mendelian loci of those complexity inherited traits. The main practical limitation to localizing QTLs seems to be the availability of suitable markers (Thoday, l96l). This limitation was remedied by the construction of complete RFLP linkage maps permitting systematic searches of an entire genome for QTLs influencing a trait (Paterson et al., l988).

In general traits influenced by several genes, the effects of any one gene are partly masked by other genes/ or by environment. Thus it is difficult to discern the effect of any one gene by merely looking at the appearance (phenotype) of the individual. Using DNA markers, QTLs can be described by their chromosomal location, dosage effect, phenotypic effects and sensitivity to the environment (Paterson *et al*., l99l). The identification and recombination of non-allelic polygenes can encourage breeders to accumulate the genes with like effect distributed in genetic materials to produce transgressive variation from which one can obtain the merit of true breeding genotypes (Xu, l989). This has been demonstrated using molecular markers in an interspecific cross of tomato (de Vicente and Tanksley, l993). Several analytical approaches have been developed to mapping QTLs with molecular markers. These include single marker analyses such as comparison of marker means, by ANOVA, regression analysis, and likelihood approach and many marker analyses using the methods of interval mapping. Lander and Botstein (l986) and Paterson *et al*. (l988) discussed new methods to map complex traits using a complete genetic maps.

As far as QTL mapping is concerned, the phenotype must be evaluated in well replicated trials in different environments. Considering this point, the use of RILs or DHs can be a good strategy for evaluating identical genotypes in different environments and

simultaneously scoring them for molecular markers (McCough, l993). The early segregating generations of a cross may not represent the optimal strategy for the complete analysis of QTL-marker association.

Mapping genes and Marker Aided Selection

The development of molecular markers promises to overcome most of previous limitations associated with morphological markers. Tight linkage of a marker to a gene can be exploited for indirect selection of traits in a breeding programme. Two prerequisites for adopting marker aided selection in breeding programmes are:1) a tightly linked marker to the gene concerned and 2) a population which is polymorphic for the marker and the gene which are in extreme linkage disequilibrium.Several aspects regarding marker-aided selection have been discussed by Beckman and Soller (l989); Stuber (l989) and Melchinger (l990).

In plant breeding, there are two distinct methods of selections are followed- one is for germplasm improvement (recurrent selection) and other for cultivar or hybrid development. These two applications are separated because recurrent selection usually is applied to random mating populations possibly at or near linkage equilibrium, whereas cultivar or hybrid development typically begins with populations derived by crossing elite inbred lines or near maximum linkage disequilibrium.

Lande and Thompson (l990) and Lande (l992) investigated the efficiency of MAS for both individual and mass selection in random mating populations. There are three approaches to applying MAS to plant breeding. l) selection markers alone with no measurement of phenotype, 2) simultaneous selection on markers and phenotype, and 3) two stage selection, the first stage involving use of markers to select among seedlings and second involving phenotypic selection among surviving individuals. The potential efficiency of MAS depends upon the heritability of the trait, the proportion of genetic variance explained by the markers, and the selection method. A major practical problem in using MAS is that recombination will reduce linkage disequilibrium between the markers and genes, thus diminishing selection effectiveness. The successful application

of MAS will require very tight linkages between marker and the trait.

Conclusion

Plant breeding efforts always include an element of chance, because numbers of genotypes and environments and consequent phenotypes to be evaluated or limitless (Wallace, l985). The molecular markers may complement the efforts to reduce the burden to a greater extent, especially in the selection of parents (based on molecular marker diversity), improving the screens for the selection of qualitative and quantitative traits, and understanding the architecture of the trait. Paterson *et al*., l99l and Lamkey and Lee, l99l discussed all the details about the molecular markers and their potentials in crop improvement. Mapping genes using DNA markers and using the DNA markers for marker assisted selection facilitate the plant breeders to speed up the process of selection in crop breeding.

References

Beckmann, J.S. and Soller, M. 1989. Genomic genetics in plant breeding in Science for plant breeding. *Proceedings of the XII Congress of EUCARPIA*. Germany. 9l-l06.

De Vicente, M.C. and Tannksley, S.D. 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. *Genetics*. l34,585-596.

Lamkey, K.R. and Lee, M. l993. Quantitative genetics. Molecular markers and plant improvement in "Focused plant improvement. *Proceedings of Tenth Australian Plant Breeding Conference, l8-23 April, Goldcoast.* 1, l04-ll5.

Lande, R. l992. MMarker assisted selection in relation to traditional methods of plant breeding. In "plant breeding in the l990s' (Edes H. R. Stalker, and J.P. Murphy) *CAB International, Wallingford, UK.* 437-451.

Martin, G.B., Williams, J.G.K. and Tanksley, S.D. l99l. Rapid identification of markers linked to Pseudomonas resistance gene in tomato by using random primers and near-isogenic lines. *Proc. Natl. Acad. Sci*. 88,2336-2340.

McCouch, S.R. l993. Progress and problems in the application of genetic maps and markers to rice improvement. *Rice Biotech Quart*., l5,43-45.

Michelmore, R.W., Paran, I. and Kesseli, R.V. l99l. Identification of markers linked to disease resistance genes by bulked segregant analysis: A rapid method to detect mamrkers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA*. 88, 9828-9832.

Muehlbauer, G.J., Specht, J.E., Thomas-Compton, M.A., Staswick, P. E. and Bernard, R. L. l988. Near isogenic lines - A potential resource in the integration of conventional and mmolecular linkage maps. *Crop Sci.* 28,729-735.

Paran, I. and MMichelmore, R.W. l993. Development of reliable PCR based markers linked to downey mildew resistance genes in lettuce. *Theor. Appl. Genet*. 85, 985-993.

Paterson, A.H., Lander, E.S., Hewitt, J.D., Paterson, S., Lincoln, S.E. and Tanksley, S.D. l988. Resolution of quantitative traits into mendelian factors by using complete linkage map of restriction fragment length polymorphism. *Nature*. 335, 721-726.

Paterson, A.H., Tanksley, S.D. and Sorells, M. E. l99l. DNA markers in plant iproveent. *Adv. Agron*. 46,39-89.

Stuber, C.W. 1989. Marker based selection for quantitative traits. in Science for Plant Breeding. *Proceedings of the XII. Congress of EUCARPIA, Germany*. 31-49.

Thoday, J.M. l96l. Location of polygenes. *Nature*.191, 368-370.

Wallace, D.H. l985. Physiological genetics of plant maturity, adaptation and yield. *Plant Breed Rev*. 3, 21-167.

Xu, Y.B. 1989. Allelism test for polygenes and its application to plant breeding. *Proceedings of the Sixth International Congress of SABRAO*. 693-696.

Association Mapping in Crop Plants

Dr. K K Vinod,
IARI, Rice Breeding and Genetics Research Centre,
Aduthurai 612101, Tanjavur District.
kkvinodh@gmail.com

Determining the genetic basis of economically important complex traits is a major goal of plant breeding, and has largely been accomplished using linkage mapping of quantitative trait loci (QTL). However, focus is now turning towards the use of association mapping (Mackay and Powell 2006), initially applied in human disease genetics. Both approaches rely on the strength of associations between genetic markers and phenotype. However, while linkage analysis searches for associations within populations developed from bi-parental crosses, association mapping utilizes historic patterns of recombination that have occurred within a sample of individuals (e.g. a collection of varieties, landraces or breeders' lines). This has the advantage of allowing existing collections to be screened for many different phenotypes, as well as taking advantage of historical phenotype data from lines thoroughly characterized during variety development. Association mapping is based on the principle that over multiple generations of recombination, correlations only with markers tightly linked to the trait of interest will remain. A best candidate for association mapping should have in its genome extensive blocks of chromatin in linkage disequilibrium (LD), providing a well-defined haplotype structure from which marker-trait associations can be identified (Otto et al 2006; Rostocks et al. 2006). However, spurious associations between genotype and trait may be detected due to the degree of structure or subdivision within the population, necessitating development of various statistical methods to account for population structure (Balding 2006).

Linkage disequilibrium (LD) is defined as a nonrandom association of alleles at separate loci located on the same chromosome (Mackay and Powell 2007). Therefore, the presence of LD is a prerequisite for association mapping where the LD extent or the physical size of LD blocks (haplotype blocks), that is chromosomal regions across which all pairs of adjacent loci are in LD (Stich 2006), determines the marker density required

for association mapping. LD can be used for a variety of purposes in crop plant genomics research. Other than the use to study marker-trait association in plants (without the use of a mapping population) followed by marker-assisted selection (MAS), another important application is to use in the population genetics studies and genetic diversity in natural populations and germplasm collections and in crop improvement programmes. Therefore combining both the uses, LD-based association mapping can be applied to germplasm bank collections, synthetic populations, and elite germplasm. Genetic association mapping or linkage disequilibrium mapping is a method that relies on linkage disequilibrium to study the relationship between phenotypic variation and genetic polymorphisms (Breseghello and Sorrells 2006).

In contrast to QTL mapping, where typically biparental crosses with contrasting genotypes are used, in the case of association studies a collection of cultivars, lines, or landraces are genotyped with densely spaced markers. In plant genetics, using a collection of cultivars has a number of advantages over the use of a bi-parental cross. Firstly, in the population a broader genetic variation in a more representative genetic background will be available. This implies that one is not limited to the marker and trait loci that happen to differ between two parents (Kraakman et al. 2006). Secondly, LD mapping may attain a higher resolution, because of the use of all meioses accumulated in the breeding history. Thirdly, historic phenotypic data on cultivars can be used to link markers to traits, without the need for new trials with special mapping populations.

Gametic phase disequilibrium

The term gametic phase disequilibrium (GPD) is used synonymously with the term "linkage disequilibrium," but GPD does not reference to linkage (as unlinked markers can still be in GPD) and emphasizes that associated alleles must co-occur in gametes. These blocks of associated alleles on a physical chromatin chunk can be known as haplotype blocks. Two alleles at distinct loci are in positive GPD if they occur together more often than predicted on the basis of their individual frequencies. This definition of association does not say anything concerning the physical position of the loci or of the alleles' joint effects on the phenotype. A statistical association between a neutral marker

allele and the phenotype occurs when marker alleles are in gametic phase disequilibrium (GPD) with alleles at a QTL.

The central problem with approaches for fine mapping using pedigree method is the limited number of meioses that have occurred and (in the case of advanced intercross lines) the cost of propagating lines to allow for a sufficient number of meioses. Further, there can be other distinct QTLs that occurs commonly in the population but were absent in the selected parents for pedigree based fine mapping. Comparing a distinct group of GPDs distributed across a natural population would therefore be a more viable approach in identifying QTLs responsible for a particular trait. Hence association mapping takes advantage of events that created association in the relatively distant past. Assuming many generations, and therefore meioses, have elapsed since these events, recombination will have removed association between a QTL and any marker not tightly linked to it. Association mapping thus allows for much finer mapping than standard bi-parental cross approaches.

Causes for gametic phase disequilibrium

In order to identify marker-trait associations, GPD or LD has to occur in the plant germplasm. LD may increase due to selection in a population, for instance when an important trait is regulated by multiple loci, or due to recent introductions of genotypes. LD may be more in self-pollinated crops, than cross pollinated. Factors contributing to the increase of LD include also small population size, inbreeding, genetic isolation between lineages, population subdivision, low recombination rate, population admixture, genetic drift and epistasis. While factors like outcrossing, high recombination rate, high mutation rate, gene conversion, etc., lead to a decrease/disruption in LD.

A variety of mechanisms generate linkage disequilibrium, and several of these can operate simultaneously. Some of the more common mechanisms are:

1. Populations expanding from a small number of founders. The haplotypes present in the founders will be more frequent than expected under equilibrium. Three cases can operate here.
    a. Genetic drift affects GPD, in which a population derived out of drift carries fewer

haplotypes than the founders.

b. A new mutation in founder, its descendants will predominantly receive the mutation and loci linked to it in the same phase. Linked marker alleles will therefore be in GPD with the mutant allele.

c. Third, an extreme case arises in the F_2 population derived from the cross of two inbred lines. The individuals derived are skewed towards a single founder.

2. In structured populations, GPD arises when allelic frequencies differ at two loci across subpopulations, irrespective of the linkage status of the loci. Admixed populations, formed by the union of previously separate populations into a single panmictic one, can be considered a case of a structured population where sub structuring has recently ceased.
3. Negative GPD will occur between loci affecting a character in populations under stabilizing or directional selection as a result of the Bulmer effect. In Bulmer effect, genetic variation is reduced over generations due to selection, and in proportion to how different the phenotype of a specific progeny's parents are in relation to the generation the parents come from.
4. Positive GPD will occur between loci affecting a character under disruptive selection.
5. When loci interact epistatically, haplotypes carrying the allelic combination favoured by selection will also be at higher-than-expected frequencies.

Estimating the gametic phase disequilibrium

Let two marker alleles, *M* and *m* having frequency of 0.6 and 0.4 are linked to a QTL with allelic frequencies of 0.5 each for *Q* and *q*. They are found in following frequencies in a population.

		QTL allele		
		Q	*q*	
Marker allele	*M*	0.4	0.2	0.6
	m	0.1	0.3	0.4
		0.5	0.5	

In the table above, the combination of alleles (or haplotype) *QM* is observed with frequency, p_{QM} = 0.4, while its predicted frequency is only $p_Q p_M$ = (0.6 x 0.5) = 0.3. The alleles *Q* and *M* are in GPD with disequilibrium coefficient, D = $p_{QM} - p_Q p_M$ = cov(*Q,M*) = 0.1. Since D can be expressed as a covariance, it can be bound its possible values by considering the case when the correlation is +/- 1, giving

$$| D | < \sigma_Q \sigma_M = [p_Q(1-p_Q)p_M(1-p_M)]^{1/2} \quad (1)$$

For a pair of biallelic loci, the expected value of the estimate of *D* is equal in magnitude irrespective of the haplotype frequencies used and can be calculated as $D = p_{QM}p_{qm} - p_{Qm}p_{qM}$. For each generation of random mating, *D* decays by a factor of (1 - *r)*, where *r* is the recombination rate between the two loci considered. Thus, after *t* generations, only *(1 -r) t* of the initial disequilibrium remains.

Therefore, LD is calculated pairwise between two polymorphic sites; using LD measures such as *D'* and r^2. The *D'* is the standardized disequilibrium coefficient which mainly measures recombinational history and is therefore useful to assess the probability of historical recombination in a given population. The r^2 is essentially the correlation between the alleles at two loci; it summarizes both recombinational and mutational history and is useful in the context of association studies. Both parameters vary in the interval from 0 to the value of 1.

To visualize or depict the extent of LD one can present a plot of LD decay which shows how LD declines with genetic (centiMorgans, cM) or physical (base pairs, bp) distance. Alternatively it is possible to construct the Disequilibrium Matrix which shows all loci in LD with corresponding probabilities. The matrix can cover the whole genome or distinct genomic locus. In general, the extent of LD varies greatly along the genome, so averages, while useful to know, may not reflect the local extent of LD. This makes the estimates of the number of markers needed more problematic. In addition, there are large variations in recombination frequency along the genome (lower near centromeres) which will affect LD in these regions.

Assessing the marker – phenotype association

Since we have data on genotypes segregating for a large number of markers, and their phenotype classes, marker – trait association can be made in two simple ways. (a) Classify the phenotypes into distinct phenotypic classes and compare the allelic frequencies across the classes and (b) group the marker genotypes and compare the phenotypes across the genotype groups. However, these methods can often end up with messed up results arising out of spurious associations, which are indistinguishable from the true ones. Thus obvious weakness of group-comparison studies is that the grouping method may result in groups that contain predominantly individuals from different subpopulations. To eliminate this weakness, family-based control methods seek case and control individuals or marker alleles within the same family.

The methods to study of marker-trait association using LD, may differ for discrete traits and quantitative traits, although sometimes quantitative traits may also be treated as discrete traits. Two procedures that have been commonly used for mapping of discrete traits (disease genes) are (i) case-control (CC) method (disease vs healthy) and (ii) transmission/ disequilibrium test (TDT) (Spielman and Ewens 1996; Allison 1997). Similar (but not identical) approaches have also been used in crop plants (Gupta et al. 2005). For quantitative traits, categorized quantitative traits are used as in CC method, and comparison of allele frequencies across groups and distinguished marker genotype groups are compared with phenotype means. Common statistical tests used are chi-square test for qualitative traits and regressoion, ANOVA and non-parametric approaches such as Kruskal-Wallis test for quantitative traits.

Transmission / disequilibrium test (TDT)

Studies on the problem of population admixture in human disease mapping, Spielman et al. (1993) developed a test called transmission / disequilibrium test (TDT) to develop unbiased association estimators. For an outbred species, the test employs family trios consisting of both parents and a progeny that belongs to one category of a dichotomous trait. One of the parents must be heterozygous and carry one copy of the focal marker allele putatively linked to the trait allele. The test consists of determining

the frequency of transmission of the focal allele to affected progeny. A chi-square or binomial test can determine whether that frequency deviates from the expectation of 0.5. Two conditions are necessary for a significant deviation: the marker allele must be both in GPD with and also linked to the allele. In the TDT, both case and control marker alleles are in effect within the same heterozygote parent. Random Mendelian segregation therefore ensures that the distribution of the TDT statistic under the null hypothesis is unaffected by population structure or selection within the pedigree (Spielman and Ewens 1996).

Extending the application of TDT to quantitative traits, Bink et al. (2000) grouped the categories of genotypes based on the quantitative trait as 'selected' and 'unselected' bringing the class into a dichotomous pattern suitable for TDT. Further considering in the standard TDT, QTL x E would lead to environmental influences on the transmission frequency of the focal marker allele from a heterozygotic parent to affected progeny. Such an effect could be detected by grouping family trios according to their environment or level of exposure to a risk factor. Heterogeneity of transmission frequency across groups would provide evidence in favor of QTL x E (Schaid 1999). Similarly, for the Monks and Kaplan test, environments would affect the magnitude of T_{MK} in the presence of QTL x E. Existence of QTL x E could then be inferred if the variance of T_{MK} across environments is significantly greater than zero. In observational studies where environments cannot be randomized across family trios, interpretation of such a result would need to be treated carefully: an association between environments and different subpopulations could also lead to heterogeneity of transmission or of T_{MK} in the absence of QTL x E.

Extending TDT for multiple markers, each linked markers are treated to be at a 'supralocus', and the TDT is applied to this supralocus (McIntyre et al. 2000; Spielman and Ewens 1996). Several methods have been developed to approach this problem, to pinpoint more precise location of QTL affecting the trait. One method uses haplotype similarities at multiple loci to infer identity by descent (IBD) probabilities with phenotype resemblance among the individuals to create cladograms, and another performs mixed model analysis to see if the variance among the groups is significant for

particular traits which detect QTL in LD with the marker haplotypes. Templeton and Sing (1993) and Templeton et al. (1987) used this cladogram that estimates the evolutionary history and relationships among haplotypes. Assuming that a mutation occurred at some point in this history on one branch of the cladogram. Haplotypes along that branch will be IBD for the mutation and distinct from haplotypes along other branches. The branches of the cladogram therefore define nested sets of haplotypes that should have related associations with the phenotype. Templeton et al. (1987) developed a nesting algorithm to group haplotypes hierarchically enabling a nested analysis of variance. This approach does not localize the mutation within the set of markers used to define haplotypes, it will increase the power to detect QTL in linkage disequilibrium with those markers.

Meuwissen and Goddard (2000) used an approach to estimate the covariance matrix among haplotype effects that does help predict QTL position within the set of markers. Starting from assumptions concerning the population history since mutation caused polymorphism at the QTL (i.e., effective population size and number of generations since mutation) the algorithm repeatedly simulates haplotype evolution and samples the probability of IBD status across specified categories of identity by state (IBS) among markers within haplotypes. The covariance among haplotype effects is then based on their IBS and its inferred relation to IBD. Since the probability function *P(IBD* | IBS) depends on QTL location within the set of markers, the assumed QTL location affects the haplotype covariance matrix. A maximum likelihood QTL position is inferred from the covariance matrix most consistent with the observed phenotypes. This approach assumes a single (monophyletic) polymorphism at the QTL while cladogram approach does not. While Meuwissen and Goddard (2000) show that the approach is, within limits, robust to population size and mutation age assumptions, applying the analysis to real (rather than simulated) data would be of interest in testing it.

Advantages and disadvantages – What to choose when?

The use of LD for mapping of QTLs for a quantitative trait is more challenging, but is also more rewarding, because it allows more precise locating of the position of a

QTL controlling the trait of interest. When comparing linkage analysis and LD mapping for QTL detection, it is revealed that linkage mapping is more useful for genome wide scan for QTLs, while LD mapping gives more precise location of an individual QTL. Therefore linkage analysis may be preferred for preliminary location of QTLs and then use LD for more precise location (Mackay 2001; Glazier et al. 2002).

Association mapping is only capable of identifying phenotypic effects of alleles with reasonably high frequency in the population under investigation. Rare alleles usually cannot be evaluated because of lack of power (not enough individuals carrying this allele). So, for such alleles classical biparental mapping can be more appropriate.

The efficiency of association mapping is significantly influenced by the population structure. The presence of population stratification and an unequal distribution of alleles facilitate mapping and identification of the underlying causes of quantitative trait variation in plants. Subgroups can result in non-functional, spurious associations. Highly significant LD between polymorphisms on different chromosomes may produce associations between a marker and a phenotype, even though the marker is not physically linked to the locus responsible for the phenotypic variation (Pritchard and Rosenberg 1999).

The complex breeding history of many important crops and the limited gene flow in most wild plants have created complex stratification within the germplasm, which complicates association studies (Sharbel et al. 2000). Association tests that do not attempt to account for the effects of population structure must be viewed with skepticism. However, recent developments in statistical methodologies make it possible to properly interpret the results of association tests. All of these methods assume that population structure has similar effects on all loci and rely on the use of independent marker loci to detect stratified populations and to correct for them (Pritchard and Rosenberg 1999).

Pritchard et al. (2000) have developed an approach that incorporates estimates of population structure directly into the association test statistic. The essential idea of the method is to decompose a sample drawn from a mixed population into several unstructured subpopulations and test the association in the homogeneous subpopulations.

The methods have been applied to association analyses in crop plants, with modified test statistics being used to deal with quantitative traits (Thornsberry et al. 2001; Beló et al. 2008). In a study of flowering time locus in maize a suite of polymorphisms in the maize *dwarf8* gene was significantly associated with variation in flowering time (Thornsberry et al. 2001). The incidence of false positives created by population structure was reduced by up to 8% as a result of the Pritchard method. Using these statistical methods in an association test allowed researchers to improve their resolution from the level of a 20-cM region to that of an individual gene.

In the other research whole genome scan association mapping was used to identify loci with major effect on oleic acid content in maize kernels, and molecular marker at about 2 kb from a fatty acid desaturase, *fad2*, was associated with the differences in the phenotype (Beló et al. 2008). The methodological advances that estimate the effects of population structure-induced linkage disequilibria should allow the use of association testing in a much wider context, enabling the use of this very powerful technique.

The other method developed by Reich and Goldstein (2001) examines the association of a moderate number of unlinked genetic markers with a given phenotype. The strength of these associations is then compared with the association of a candidate gene.

Nowadays there exists a handful of published software to assess the association of marker loci with traits. The most commonly used statistics include logistic regression with the possibility of structured associations implemented in TASSEL General Linear Model (Yu and Buckler 2006; TASSEL: http://www.maizegenetics.net), a multiple regression model combined with the estimates for the false discovery rate suggested by Kraakman et al. (2006), and an unified mixed-model approach described by Yu et al. (2006) and implemented in TASSEL Mixed Linear Model or in SAS v9.1.2 (Ehrenreich et al. 2007).

Factors affecting precision of association mapping

Experiences show that following factors may affect the precision of association mapping.

a. Magnitude of the GDP

b. Effective size of the QTL allele

c. Mod of gene action (dominance, additive or multiplicative)

Practical approach to association mapping

As in the case of linkage based mapping, precision must be maintained with approaches of genotyping and phenotyping in plants.

Genotyping

There are many types of markers that can be used for this, including AFLP and single sequence repeats (SSRs, also known as microsatellite markers). AFLP markers are easily obtained in almost any organism, even for those lacking previously existing genomics data. However, AFLP markers are almost exclusively dominant, that is the heterozygous genotype cannot be distinguished from one of the homozygous genotypes, and this introduces a number of problems when using AFLP markers for estimating, for instance, population structure or for use directly in mapping. SSR markers, on the other hand, are usually highly polymorphic but require a great deal of work to isolate and are rarely transferable between anything but the most closely related species. The high variability of SSR markers, combined with the availability of semi-automatic detection methods, have, until recently, made them the markers of choice for use in estimating population structure or pairwise relatedness among individuals.

The development of next-generation sequencing technologies has allowed for unprecedented genotyping capabilities, even in organisms that have traditionally not been considered. The current next generation sequencing technologies are capable of analyzing anywhere from hundreds of thousands to tens of millions of DNA molecules in parallel compared with hundreds at a time which is the maximum throughput of most traditional sequencing instruments. Next-generation sequencing technologies allows for rapid

identification of a large number of genetic markers, mainly single nucleotide polymorphisms (SNPs). SNP markers have both a higher genome density and a lower mutation rate than SSR markers and they are also more easily amenable to high-throughput genotyping in multiplex or microarray format. The mutational processes underlying SNP variation is well-understood whereas the mutational processes of other types of markers, such as SSRs are poorly understood and this sometimes hampers analyses using such markers. The vast majority of SNPs are bi-allelic and the information content per marker is therefore much lower than in SSR markers. This, however, is more than compensated for by the fact that they are more widely distributed across the genome in most organisms. SNP markers are therefore rapidly becoming the markers of choice for most association mapping studies in both model and non-model plant species (Hall et al. 2010).

One issue that has been receiving an increasing interest is how the selection of SNPs to include in an association study can potentially bias the results. For instance, SNP discovery panels are often small, suggesting that low-frequency mutations are more likely to go undetected. This will bias the frequency spectrum of the identified mutations compared with what would be obtained from the full sample, with relatively more SNPs occurring at intermediate frequencies. The ascertainment bias introduced in the SNP selection process have important consequences for any inferences that are drawn from the data; for association mapping the most detrimental effect is an over-sampling of mutations at intermediate frequencies which results in lower levels of linkage disequilibrium (LD) than if SNPs were selected completely at random. The effect of ascertainment bias on the power of association studies is more complex, and largely depends on whether low or intermediate frequency are assumed to have a larger effect on the trait of interest.

Another important problem to be aware of in association mapping is the genotyping error rate. While state-of-the-art SNP scoring methods are usually quite robust, the rate of genotyping errors can vary a lot between different SNPs even when scored on a single chip. This is important to remember since even low error rates (around

3% or less) can have dramatic consequences for the accuracy of estimates of LD and hence also for association mapping.

Candidate genes versus whole genome scans

The most important aspect when deciding between a candidate gene approach and a whole-genome study for association mapping is the extent of LD in the organism of interest, because the extent of LD determines not only the mapping resolution that can be achieved, but also the numbers of markers that are needed for an adequate coverage of the genome in a genome-wide study. When considering the extent of LD one should preferably also account for variation of recombination rates across the genome, although this may be hard to implement in organisms where regional variation in LD is poorly documented. In species where LD extends over long physical distances, relatively few markers are needed to ensure adequate genome coverage; for example the extensive LD seen in species like *Arabidopsis thaliana* or in inbred lines of barley, where LD can extend for tens or even hundreds of kilo base pairs, allow for genome-wide association mapping with a relatively low number of evenly spaced SNPs markers. However, in cross pollinated plants, such as maize and many forest trees, LD only extends a few hundred base pairs at the most and adequate genome-wide coverage would require several million SNPs.

The alternative approach is to perform a candidate gene-based association study. A candidate- gene association mapping study is more hypothesis-driven than a genome-wide study, since association mapping is restricted to relevant candidates genes thought to be involved in controlling the trait of interest. The selection of candidates is not straightforward, but choices can be based on relevant information obtained from, for instance, genetic, biochemical, or physiology studies in both model and non-model plant species. Candidate gene selection is usually quite straightforward when restricted to well characterized developmental pathways, like the flowering pathways in *Arabidopsis* and other plants, or to traits with a well-understood biochemical basis, such as the starch-synthesis pathway in maize. Candidate gene studies are less demanding in terms of the number of markers that are required and many candidate gene association studies have

successfully been completed using tens to hundreds of markers in mapping populations consisting of a few hundred individuals. However, it is important to remember that a candidate gene approach is limited by the choice of candidate genes that are identified and hence always runs the risk of missing out on identifying causal mutations that are located in non-identified candidate genes. In addition, candidate genes are often initially discovered from loss-of-function mutations in inbred lab strains and it is not clear how well such mutations describe the variation that actually underlie quantitative trait variation in natural populations.

Phenotyping

As the cost of genotyping is rapidly declining, a greater fraction of the budget of any association mapping project will be spent on phenotyping. In fact, while the importance of accurate identification and scoring of genotypes have received quite a deal of, the effects of phenotyping have yet to be evaluated in any greater detail. It has been shown, however, that increasing the number of individuals phenotyped is far more efficient than increasing the number of SNPs for increasing the power in association studies. Also, several new experimental designs are actively being developed that combine the best aspects of traditional QTL mapping and association mapping (e.g. nested association mapping, Yu et al. 2008).

A typical association mapping study usually involves a diverse set of accessions, and phenotypic scoring with adequate accuracy can be both costly and time consuming. Replication of individual accessions within a site is usually needed to increase precision in phenotypic measurements, by eliminating environmentally induced noise and measurement errors. Data on replicates of each accession can then be combined to produce an estimate of the 'mean' phenotype of the accession which is less influenced by environment or measurement errors. One example of such an approach is the estimation of breeding values which is common practice in quantitative genetics and breeding. These breeding values are used as dependent traits in an association analysis in an attempt to dissect the genetic basis of the trait in question (Stich et al. 2008).

An additional benefit of replication can be achieved if the entire association mapping collection is replicated across multiple environments. Such a design can provide important information on the robustness of positive associations across environments and on the importance of genotype by environment interactions in shaping allelic contributions to the trait of interest.

Controlling for population structure

One of the main hurdles for using association mapping to dissect the genetic architecture of complex traits in plants is the risk of incurring false positives due to population structure (Pritchard et al. 2000b; Zhao et al. 2007). The problem of population structure arises because any phenotypic trait that is also correlated with the underlying population structure at neutral loci will show an inflated number of positive associations. The problem of population structure is well known and many methods have, not surprisingly, been developed to deal with this problem.

One of the first methods proposed was the method of 'genomic control' (GC) developed by Devlin and Roeder (1999). The rationale for GC is to estimate association using a large number of putative neutral markers or markers not thought to be involved in controlling the trait of interest. The distribution of the test statistic of interest is then calculated from these associations and a critical value corresponding to the desired Type I error rate is chosen from this distribution. While GC is straightforward to perform computationally, it requires a large number of control loci to accurately capture the extent of variation in population structure across the genome of an organism. Furthermore, in some situations it is possible for GC to 'overcorrect' for population structure effects resulting in a loss of power to detect true associations.

Another method that is commonly used to control for population structure is structured associations (SA). The idea of SA builds on the general linear model (GLM) method of Pritchard et al. (2000b) who infer details of population structure and the ancestry of sampled individuals using a set of unlinked genetic markers. This information is then used to identify populations within which mating is random. Markers are then tested for associations within these sub-populations identified by the genetic markers.

The most recent and most promising approach, for correcting the spurious effects of population structure is the mixed linear model (MLM) approach outlined by Yu et al. (2006). MLM use information on both population structure and more cryptic relatedness among members of an association study to correct for the spurious effects of populations structure and relatedness. These two types of population structure are incorporated into a matrix of population effects (Q) and a matrix describing the relative kinship of individuals in a sample (K) and a model is then fitted using the MLM framework. The Q matrix consists of one or more vectors describing the underlying population structure and this matrix can be estimated in several ways. For example, one common approach is to use the method implemented in STRUCTURE (Pritchard et al. 2000a) or by using principle component analysis (PCA) of the complete genotype data (Patterson et al. 2006). Using PCA to estimate population structure is especially appealing since it is far less computationally demanding than analyses based on STRUCTURE (Patterson et al 2006: Price et al. 2006). A similar approach to PCA is to use nonmetric multidimensional scaling (nMDS, Zhu and Yu 2009) which have been shown to reduce the false positive rate compared to other methods in structured populations.

The kinship matrix (K), on the other hand, can be estimated from pedigree data or, for non-model species where pedigree information is usually lacking, using relative kinship coefficients estimated using genetic marker data. The strength of the MLM approach is that it handles and performs well under many types of population structure. For instance, in a genome-wide association study in *Arabidopsis thaliana*, the MLM provided the most accurate control of the false-positive rate among the methods tested, despite a very complex sub-structuring of the association population (Zhao et al. 2007).

The original intent of the Q and K matrices is to capture different types of population structure (Yu et al. 2006) and several studies have found that including either the Q or the K matrix alone is not sufficient to control for all aspects of the underlying population structure of the data. However, the relative utility of the two matrices depends on the actual pattern of the underlying population structure. Both STRUCTURE and the PCA-based analyses have problems identifying low levels of population structure when a low to moderate number of markers are used. Patterson et al. (2006) even defined a

minimum study design that is needed to effectively evaluate population structure and showed that for a given design there exists a minimum level of population structure that can be detected. For example, a STRUCTURE-based analysis failed to identify any obvious signs of population structure in European aspen (*Populus tremula*), despite evidence for significant population structure and isolation-by-distance based on population groupings chosen *a priori* (Hall et al. 2007). However, the same set of markers, when used to estimate the K for the sampled trees, provided a reasonable control of the underlying weak, but nevertheless significant, isolation by distance (Ingvarsson et al. 2008).

Replication and validation

As the number of studies documenting alleles showing significant associations with quantitative trait variation, there is an increasing need to replicate findings and to validate estimates of allelic effects. Replication of genotype phenotype association are crucial for separating true from false positives and to provide less biased estimates of allelic effect sizes. However, failure to replicate a previously documented association can occur because of a large number of issues, both in the initial and the replication study, including factors like difficulties in replicating the environment, small sample size, poor study design or lack of rigorous phenotype scoring (Manolio et al. 2009). The literature on association mapping in plants does, however, include a few cases where associations have been replicated in independent mapping experiments. For example, Thornsberry et al. (2001) found that mutations in the gene *Dwarf 8* affect the quantitative variation of flowering time and plant height in maize (*Zea mays*). This association has subsequently been verified in a larger maize association mapping population containing a different set of maize inbred lines (Camus-Kulandaivelu et al. (2006). Finally, it is worth pointing out that verification of genotype–phenotype associations does not necessarily have to come from replicate association studies, but can include validation of biological function through transgenic experiments and other molecular biology techniques (Koornneef et al. 2004).

Another concern is that allelic effects of previously documented associations usually decline in replication studies. This phenomenon is known as the 'Beavis effect' (Beavis 1998) in the QTL mapping literature and occurs because significant associations are reported only when test statistics exceed a predetermined critical threshold. The estimated effects of detected associations are therefore sampled from a truncated distribution, and the weaker the initial effect the more serious this overestimation is (Rockman 2008). The Beavis effect has also been shown to occur in association mapping studies. The Beavis effect is known to be weaker when the mapping population used in the experiment is larger, hence careful consideration of the power of the prospective association study should be taken early on in the experiment, so that things like the Beavis effect can be minimized or eliminated.

Conclusions

With the rapidly dropping costs of modern sequencing and genotyping, generation of genotype data is no longer the limiting factor for most studies. This has resulted in a need for new refined statistical methods for association analysis that cover entire genomes and the greatest costs are utilized towards rigorous phenotyping instead of generation of genotypic data. Establishing specialized common gardens to minimize environmental influence and possible epigenetic effects are being practiced. Individual alleles or QTLs identified in association studies usually explain only a few percent of the variation in traits studied and even when many loci associated to a trait are taken into account the proportion of variation explained is usually far below than prediction-based heritabilities of the traits, a phenomenon highlighted in the human-genetics community as the 'missing heritability problem'. The problem of 'missing heritability' has several likely causes that are poorly accounted for in current association mapping studies, such as low-frequency alleles of large effect, allelic interactions (i.e. epistasis), copy number variation and possible epigenetic effects (Manolio et al. 2009). As more and more putative causative alleles are identified, it becomes increasingly necessary for methods that can deal with associations across gene networks of interacting genes and across developmental pathways.

An additional question that should be addressed is about how the effects of individual alleles vary across different environments. Given the ubiquity of genotype environment for many traits in plants (Lynch and Walsh 1998), to what degree QTL effects vary across environments has important implications, e.g. the utility of QTLs in breeding applications.

References

Lynch, M., Walsh, B, 1998. Genetics and Analysis of Quantitative Traits. *Sinauer Associates,* Sunderland, MA.

Allison, D.B. 1997. Transmission disequilibrium tests for quantitative traits. *Am J Hum Genet*. 60,676-690.

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. *Nat Genet* .7,781-791.

Beavis, W.D. 1998. QTL analyses: power, precision, and accuracy. *Paterson AH (ed). Molecular Dissection of ComplexTraits* .CRC Press, New York. 145–162.

Beló, A., Zheng, P., Luck, S., Shen, B., Meyer, D.J., Li, B., Tingey, S., Rafalski, A. 2008. Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Mol Genet Genomics.* 279,1-10.

Bink, M.C., Te Pas, M.F.W., Harders, F.L., Janss, L.L.G. 2000.A transmission/disequilibrium test approach to screen for quantitative trait loci in two selected lines of Large White pigs. *Genet Res*. 75, 115-121.

Breseghello, F., Sorrells, M.S. 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics.* 172,1165-1177.

Camus-Kulandaivelu, L., Veyrieras, J.B., Madur, D. et al. 2006. Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. *Genetics*. 172, 2449–2463.

Ehrenreich, I.M., Stafford, P.A., Purugganan, M.D. 2007. The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping. *Genetics.* 176, 1223-1236.

Glazier, A.M., Nadeau, J.H., Aitman, T.J. 2002 Finding genes that underlie complex traits. *Science.* 298, 2345-2349.

Gupta, P.K., Rustgi, S., Kulwal, P.L. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol Biol.* 57, 461-485.

Hall, D., Luquez, V., St. Onge K.R., et al. 2007. Adaptive population differentiation in bud phenology across a latitudinal gradient in European aspen (*Populus tremula*, L., Salicaeae): a comparison of neutral markers, candidate genes and quantitative traits. *Evolution,* 61, 2849–2860.

Hall, D., Tegström, C., Ingvarsson, P.K. 2010. Using association mapping to dissect the genetic basis of complex traits in plants. *Brief Funct Genom.* 9,157-165.

Ingvarsson, P.K., Garcia, M.V., Luquez, V, et al. 2008 Nucleotide polymorphism and phenotypic associations within and around the phytochromeB2 locus in European aspen (*Populus tremula*, Salicaceae). *Genetics*. 178, 2217–2226.

Koornneef, M., Alonso-Blanco, C., Vreugdenhil, D., 2004. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu Rev Plant Biol* .55,141–172.

Kraakman, A.T.W., Martínez, F., Mussiraliev, B., Van Eeuwijk, F.A., Niks, R.E. 2006 Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. *Mol Breed* .17,41-58.

Mackay, I., Powell, W. 2006. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 12,57-63.

Mackay, T.F.C. 2001. The genetic architecture of quantitative traits. *Annu Rev Genet* 33,303-339.

Manolio, T.A., Collins, F.S., Cox, N.J., et al. 2009. Finding the missing heritability of complex diseases. *Nature* .461,747–753.

McIntyre, L.M., Martin, E.R., Simonsen, K.L., Kaplan, N.L. 2000 Circumventing multiple testing: A multilocus Monte Carlo approach to testing for association. *Genet Epidem* 19, 18-29.

Meuwissen, T.H.E., Goddard, M.E. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*. 155, 421-430.

Otto, L.V.M., Ganal, M.W., Röder, M.S. 2006. Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet* . 7, 6.

Patterson. N., Price, A.L., Reich, D 2006. Population structure and eigen analysis. *PLoS Genet.* 2,e190.

Price, A.L., Patterson, N.J., Plenge, R.M., et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* .38, 904–909.

Pritchard, J.K., Rosenberg, N.A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet.* 65,220-228.

Pritchard, J.K., Stephens, M., Donnelly, P. 2000a. Inference of population structure using multilocus genotype data. *Genetics*. 155,945–959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P 2000b. Association mapping in structured populations. *Am J Hum Genet.* 37,170-181.

Reich, D.E., Goldstein, D.B. 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* .20,4-16.

Rockman, M.V. 2008. Reverse engineering the genotype phenotype map with natural genetic variation. *Nature*. 456,738–744.

Rostocks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P.R., Roose, M.L., Svensson, J.T., Stein, N., Varshney, R.K., Marshall, D.F., Graner, A., Close, T.J., Waugh, R. 2006 Recent history of artificial outcrossing facilitates whole genome association mapping in elite crop varieties. *Proc Natl Acad Sci USA* .103,18656-18661.

Schaid, D.J. 1999. Case-parents design for gene-environment interaction. *Genet Epidem* 16, 261-273.

Spielman, R.S., Ewens, W.J. 1996. The TDT and other family based tests for linkage disequilibrium and association. *Am J Hum Genet* .59,983-989.

Spielman, R.S., McGinnis, R.E., Ewens, W.J. 1993. Transmission test for linage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* .52,506-516.

Stich, B. 2006. A new test for family-based association mapping with inbred lines from plant breeding programs. *Theor Appl Genet* 113,1121-1130.
Stich, B., Möhring, J., Piepho, H.P., et al. 2008. Comparison of mixed-model approaches for association mapping. *Genetics.* 178,1745–1754.

Templeton, A.R., Boerwinkle, E., Sing, C.F. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*.117, 343-351.

Templeton, A.R., Sing, C.F. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics.* 134,659-669.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S. IV 2001 *Dwarf8* polymorphisms associate with variation in flowering time. *Nat Genet.* 28,286-289.

Yu, J., Buckler, E.S. 2006. Genetic association mapping and genome organization of maize. *Curr Opin Biotechn.* 17,155-160.

Yu, J., Holland, J.B., McMullen, M.D., Buckler, E.S. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics.* 178,539–551.

Yu, J., Pressoir, G., Briggs, W.H., et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 38, 203–208.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 38,203-208.

Zhao, K., Aranzana, M.J., Kim, S., et al. 2007. An Arabidopsis example of association mapping in structured samples. *PLoS Genetics.* 3,e4.

Zhu, C., Yu, J. 2009. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics,* 182,875–888.

Whole Genome Sequencing and Annotation - An overview

Dr. L.Arul, Associate Professor,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
arulsra@gmail.com

The "*Genome*" is universally defined as the total nuclear DNA content of a haploid cell or half the DNA content of a diploid cell of any given organism. Sequencing genome(s) is regarded the way for an ultimate characterization of an organism. The first genome to be sequenced was a small phage Φ-X174 of size 5.38 Kb in 1977, only after the advent of the Sanger's dideoxy method of DNA sequencing. The next major happening was the sequencing of the first bacterial genome, *Haemophilus influenzae* of 1.8 Mb in size during 1995. These were however remarkable achievements in the field of genomics which paved way for the successful sequencing of hundreds of prokaryotic and a few eukaryotic genomes so far.

Genome sequencing methods

Primarily there are two different sequencing strategies employed in case of eukaryotic genomes, the clone by clone or map based approach and whole genome shotgun sequencing (WGS). The clone by clone method, the first to be employed in human genome studies, is slow but sure. This is also referred to as the map-based method, evolved from procedures developed by a number of researchers during the late 1980s and 90s. Saturated physical maps formed the very basis for this approach. Once the genome is fragmented it becomes necessary to assemble the cloned fragments in the same linear order as found in the chromosomes from where they were derived. Physical markers which are unique DNA sequences were used for identifying the overlaps and assist the genome reconstruction process. Originally, physical maps were seen as essential tools to facilitate the sequencing of complete genomes. However, it is now possible to shotgun sequence an entire genome without the existence of a clone based map for even larger genomes. The technique of whole genome shotgun sequencing is much faster and was developed by J. Craig Venter. In this method, the entire genomic sequence is fragmented into pieces of clonable length. After amplification in the

bacterium *E. coli*, the fragments are sequenced and assembled using powerful computer algorithms. A list of select genome sequencing terminologies is given in the box 1.

Box1. Selected genomic terminologies

- **WGS** – Whole genome sequencing
- **Read** – the length of sequence generated by the sequencer
- **Contigs** – (*contiguous)* overlapping DNA fragments (sequences)
- **Scaffolds** –a series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence
- **Mb** – mega bases or million bases (10^6)
- **Coverage** – the average number of times a genomic segment is represented in a collection of clones or sequence reads
- **Minimal tiling path** - a minimal set of over-lapping clones that together provides a complete coverage across a genomic region

Concept of model plant genomes

Model organisms (*Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae*) provide genetic and molecular insights into the biology of more complex species. The effort to determine the nucleotide sequence of a plant genome has issues in the light of the genetic makeup of the plants. The range of plant genome size is very large extending from approximately the same size as the genome of many small animals through more than five times as large as the human genome for many of the domesticated crops to almost forty times as large as the human genome for some ornamental flowers. Hence, the plant biologists started resorting to the concept of model organisms characterized by shared features such as being diploid and appropriate for genetic analysis, being amenable to genetic transformation, having a (relatively) small genome and a short growth cycle, having commonly available tools and resources, and being the focus of research by a large scientific community. The species now used as model organisms for mono- and dicotyledonous plants are rice (*Oryza sativa*) and *Arabidopsis* (*Arabidopsis thaliana*) respectively.

Arabidopsis, a small plant belonging to the family Cruciferaceae of little agricultural use, it sets seed in only 6 weeks from planting, has a small genome of 120

Megabases (Mb) and only five chromosomes. There are extensive tools available for its genomic analysis, whole genome sequence, Expressed Sequence Tags (ESTs) collections, characterized mutants and large populations mutagenized with insertion elements (transposons or the T-DNA of *Agrobacterium*). *Arabidopsis* can be genetically transformed on a large scale with *Agrobacterium tumefaciens* and biolistics. Other tools available for this model plant are saturated genetic and physical maps. As a result, this was the first plant and third multicellular organism to be sequenced. The Arabidopsis Genome Initiative in 2000 came out with the complete sequence of *Arabidopsis* in *Nature* under the title "*Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*". The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *Caenorhabditis elegans,* the other sequenced multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

Unlike *Arabidopsis*, rice is one of the world's most important cereals. It is considered a model crop because it has a relatively small genome size compared with other cereals, vast germplasm collection. Rice also has very saturated genetic maps, physical maps, whole genome sequences, as well as EST collections pooled from different tissues and developmental stages. It has 12 chromosomes, a genome size of 420 Mb, and like *Arabidopsis*, it can be transformed through biolistics and *A. tumefaciens*. The scientific value of rice is further enhanced with the elucidation of genome sequence of the two major subspecies of cultivated rice, *japonica* and *indica*. The sequence of the *japonica* cultivar Nipponbare was completed by a consortium of ten countries, which comprised the International Rice Genome Sequencing Project (IRGSP). It was a publicly

funded program established in 1997 with members from Japan, United States of America, China, Taiwan, Korea, India, Thailand, France, Brazil, and the United Kingdom. The IRGSP adopts the clone-by-clone shotgun sequencing strategy so that each sequenced clone can be associated with a specific position on the genetic map and adheres to the policy of immediate release of the sequence data to the public domain. In December 2004, the IRGSP published the sequencing of the rice genome in *Nature* under the title "*The map-based sequence of the rice genome*". The high-quality and map-based sequence of the entire genome is now available in public databases. Further, the sequencing of *indica* genome was derived from whole-genome shotgun sequencing approach in 2005. These genome sequences are invaluable resources not only in understanding the structure and function of the rice plant itself but also in deciphering the organization of other cereal genomes, which share appreciable degree of synteny with rice. Besides, crop plants not only have economic significance, but also comprise important botanical models for evolution and development. This is reflected by the recent increase in the percentage of publicly available sequence data that are derived from angiosperms (Table 1 shows a list of plants for which the complete genome sequence is currently available).

Table 1: List of completely sequenced plant genomes

Sl. No	Plant	Chr. No (n)	Size (Mb)
1	*Arabidopsis thaliana* (thale cress)	5	120
2	*Oryza sativa* (rice)	12	390
3	*Glycine max* (soybean)	20	950
4	*Medicago truncatula* (barrel medic)	8	500
5	*Populus trichocarpa* (black cottonwood)	19	480
6	*Sorghum bicolor*	10	690
7	*Vitis vinifera* (wine grape)	19	500
8	*Zea mays* (corn)	10	2300

Steps immediately after sequencing

The early steps of the genome assembly and annotation involve many computationally intensive processes, including elimination of contaminants, masking the repetitive sequences and aligning each genomic sequence to the other genomic sequences, mRNAs, and Expressed Sequence Tags (ESTs).

Removing contaminants

Draft quality HTGS's sometimes contain segments of sequence derived from foreign sources, most commonly the cloning vector or bacterial host. Finished sequences are usually, but not always, free of such contaminants. Any foreign segments are removed from draft-quality sequence or masked in finished sequence to prevent them from participating in alignments. A BLAST against "UniVec" database helps in the removal of the contaminants.

Masking of repetitive sequences

Sequences that occur in many copies in the genome will align to many different clones. Such repetitive sequences include interspersed repeats (SINEs, LINEs, LTR elements, and DNA transposons), satellite sequences, and low-complexity sequences. Matches between repetitive sequences on unrelated clones make it difficult to identify alignments that indicate a genuine overlap between clones. To eliminate the confusing matches that are based only on repetitive sequences, the genomic sequences are run through a program called "RepeatMasker" to identify known repeats. Repeats are masked by converting the sequence to lowercase letters so that they do not initiate alignments.

Chromosome assignment

To improve assembly of the genomic sequences, the input genomic sequences are assigned to a specific chromosome before attempting to merge the sequences. Genomic sequences that appear on any of the chromosome tiling paths are automatically assigned to the designated chromosome. Other genomic sequences are assigned to a chromosome based on: (*a*) annotation on the submitted GenBank record; (*b*) the presence of multiple STS markers that have been mapped to the same chromosome; (*c*) fluorescence *in situ* hybridization (FISH) mapping. If there is no assignment, or the assignments are

conflicted, the sequences are treated as unassigned and assembled without constraint by chromosome.

Genome assembly

The input genomic sequences are assembled into a series of genomic sequence contigs. These are then ordered, oriented with respect to each other, and placed along each chromosome with appropriately sized gaps inserted between adjacent contigs. The resulting genome assembly thus consists of a set of genomic sequence contigs and a specification for how to arrange the sequence contigs along each chromosome.

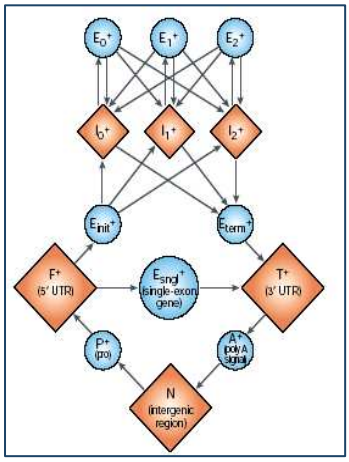
Genome annotation

Whole genome sequencing of plants and animals is churning out huge volumes of sequence data which pose to be highly incomprehensible. Understanding the genome constitution, growth and development, evolution and ultimately every other relevant issues are presently being addressed by a combination of experimental and bioinformatics tools. As a result of this synergy, there has been a tremendous progress in some of the areas which includes: *gene finding and genome annotation*, comparative genomics, allele mining, protein-protein interactions, structure-function relationships, *in silico* dissection of quantitative trait loci, global transcript profiling, deciphering metabolic and signalling pathway, drug discovery, diversity analysis and phylogeny.

Gene finding typically refers to the area of computational biology that is concerned with algorithmically identifying stretches of sequence that are biologically functional. This especially includes protein-coding genes, but may also include other functional elements such as RNA genes and regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced. In the genomes of prokaryotes, genes have specific and relatively well-understood promoter sequences (signals), such as the Pribnow box and transcription factor binding sites, which are easy to systematically identify. Also, the sequence coding for a protein occurs as one contiguous open reading frame (ORF), which is typically many hundreds or thousands of base pairs long. Furthermore, protein-coding DNA has certain periodicities and other statistical properties that are easy to detect in sequence of this length. These characteristics make prokaryotic gene finding relatively

straightforward, and well designed systems are able to achieve high levels of accuracy. However, gene finding in eukaryotes, especially complex organisms like humans, is considerably more challenging for several reasons. First, the promoter and other regulatory signals in these genomes are more complex and less well-understood than in prokaryotes, making them more difficult to reliably recognize. Two classic examples of signals identified by eukaryotic gene finders are CpG islands and binding sites for a poly(A) tail. Second, splicing mechanisms employed by eukaryotic cells mean that a particular protein-coding sequence in the genome is divided into several parts (exons), separated by non-coding sequences (introns). (Splice sites are themselves another signal that eukaryotic gene finders are often designed to identify.) A typical protein-coding gene in humans might be divided into a dozen exons, each less than two hundred base pairs in length, and some as short as twenty to thirty. It is therefore much more difficult to detect periodicities and other known content properties of protein-coding DNA in eukaryotes. Gene finder programs for both prokaryotic and eukaryotic genomes typically use statistical models, such as Hidden Markov Models, which combines biological information from a variety of different signal and content features. The GLIMMER system is a widely used and highly accurate gene finder for prokaryotes. Eukaryotic *ab initio* gene finders, by comparison, have also achieved a reasonable level of success, notable example is GENSCAN (Fig 1).

Fig 1: GENSCAN model indicating states and transition, states are parameterized with signal and content features



Proteomics – A Method for Large Scale Identification of Complex Proteins

Dr. N.Senthil and Dr. M.Raveendran, Associate Professors,
Genomics and Proteomics Laboratory
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
senthil_natesan@yahoo.com & raveendrantnau@gmail.com

Introduction

Proteins are the work-horses of the cell and have important functions in both normal and abnormal states. In order to understand how proteins interact and regulate various cellular processes, it is important to understand their expression behavior under a wide range of experimental conditions. Unlike the genome which contains a fixed number of genes, the levels of protein within the cells are highly dynamic. The term "proteomics" was first coined in 1995 and was defined as the large-scale characterization of the entire protein complement of a cell line, tissue, or organism. The goal of proteomics is to obtain a more global and integrated view of biology by studying all the proteins of a cell rather than each one individually. The aim of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional (3-D) map of the cell indicating where proteins are located. These ambitious goals will certainly require the involvement of a large number of different disciplines such as molecular biology, biochemistry, and bioinformatics.

In the quest to characterize the proteome of a given cell or organism, it should be remembered that the proteome is dynamic. The proteome of a cell will reflect the immediate environment in which it is studied. In response to internal or external cues, proteins can be modified by posttranslational modifications, undergo translocations within the cell, or be synthesized or degraded. Thus, examination of the proteome of a cell is like taking a "snapshot" of the protein environment at any given time. Considering all the possibilities, it is likely that any given genome can potentially give rise to an infinite number of proteomes.

Proteomics Origins

The first protein studies that can be called proteomics began in 1975 with the introduction of the two-dimensional gel by O'Farrell, Klose, and Scheele, who began mapping proteins from *Escherichia coli*, mouse, and guinea pig, respectively. The first major technology to emerge for the identification of proteins was the sequencing of proteins by Edman degradation . A major breakthrough was the development of microsequencing techniques for electroblotted proteins. This technique was used for the identification of proteins from 2-D gels to create the first 2-D databases. One of the most important developments in protein identification has been the development of MS technology. In the last decade, the sensitivity of analysis and accuracy of results for protein identification by MS have increased by several orders of magnitude. It is now estimated that proteins in the femtomolar range can be identified in gels. Because MS is more sensitive, can tolerate protein mixtures, and is amenable to high-throughput operations, it has essentially replaced Edman sequencing as the protein identification tool of choice.

Why Proteomics?

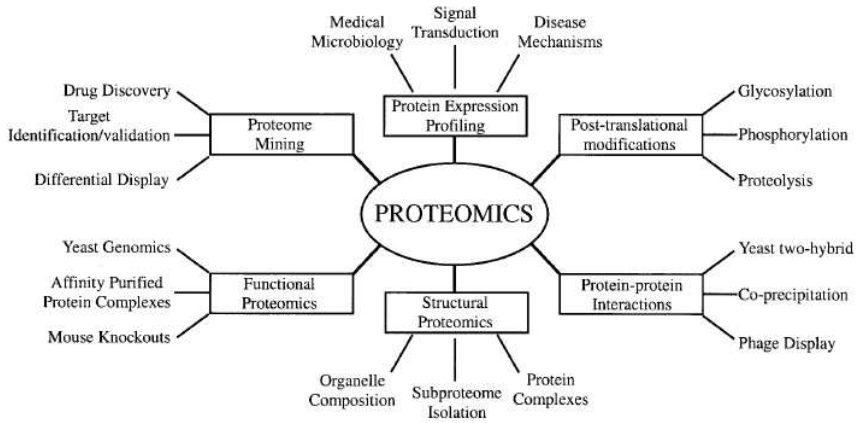
Many types of information cannot be obtained from the study of genes alone. For example, proteins, not genes, are responsible for the phenotypes of cells. It is impossible to elucidate mechanisms of disease, aging, and effects of the environment solely by studying the genome. Only through the study of proteins can protein modifications be characterized and the targets of drugs identified.

Types of Proteomics

Protein expression proteomics: The quantitative study of protein expression between samples that differ by some variable is known as expression proteomics. In this approach, protein expression of the entire proteome or of subproteomes between samples can be compared. Information from this approach can identify novel proteins in signal transduction or identify disease-specific proteins.

Structural proteomics: Proteomics studies whose goal is to map out the structure of protein complexes or the proteins present in a specific cellular organelle are known as "cell map" or structural proteomics. Structural proteomics attempts to identify all the proteins within a protein complex or organelle, determine where they are located, and characterize all protein-protein interactions.

Functional proteomics: "Functional proteomics" is a broad term for many specific, directed proteomics approaches. In some cases, specific subproteomes are isolated by affinity chromatography for further analysis. This could include the isolation of protein complexes or the use of protein ligands to isolate specific types of proteins. This approach allows a selected group of proteins to be studied and characterized and can provide important information about protein signaling, disease mechanisms or protein-drug interactions.



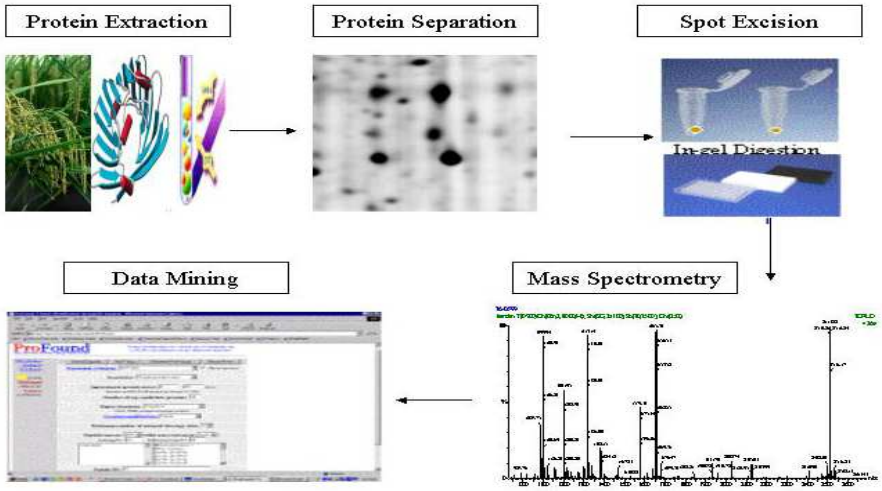
Protein analysis

By the very definition of proteomics, it is inevitable that complex protein mixtures will be encountered. Therefore, methods must exist to resolve these protein mixtures into their individual components so that the proteins can be visualized, identified, and characterized. The predominant technology for protein separation and isolation is polyacrylamide gel electrophoresis. Since its inception some 32 years ago, protein electrophoresis still remains the most effective way to resolve a complex mixture of proteins. In many applications, it is at this stage where the bottleneck occurs. This is because 1- or 2-DE is a slow, tedious procedure that is not easily automated. However,

until something replaces this methodology, it will remain an essential component of proteomics.

One- and two-dimensional gel electrophoresis.

For many proteomics applications, 1-DE is the method of choice to resolve protein mixtures. In 1-DE, proteins are separated on the basis of molecular mass. Because proteins are solubilized in sodium dodecyl sulfate (SDS), protein solubility is rarely a problem. Moreover, 1-DE is simple to perform, is reproducible, and can be used to resolve proteins with molecular masses of 10 to 300 kDa. The most common application of 1-DE is the characterization of proteins after some form of protein purification. This is because of the limited resolving power of a 1-D gel. If a more complex protein mixture such as a crude cell lysate is encountered, then 2-DE can be used. In 2-DE, proteins are separated by two distinct properties. They are resolved according to their net charge in the first dimension and according to their molecular mass in the second dimension. The combination of these two techniques produces resolution far exceeding that obtained in 1-DE. One of the greatest strengths of 2-DE is the ability to resolve proteins that have undergone some form of posttranslational modification. This resolution is possible in 2-DE because many types of protein modifications confer a difference in charge as well as a change in mass on the protein. One such example is protein phosphorylation. Frequently, the phosphorylated form of a protein can be resolved from the nonphosphorylated form by 2-DE.



Alternatives to electrophoresis.

The limitations of 2-DE have inspired a number of approaches to bypass protein gel electrophoresis. One approach is to convert an entire protein mixture to peptides (usually by digestion with trypsin) and then purify the peptides before subjecting them to analysis by MS. Various methods for peptide purification have been devised, including liquid chromatography, capillary electrophoresis and a combination of techniques such as multidimensional protein identification or cation-exchange chromatography and reverse-phase (RP) chromatography. The advantage of these methods is that because a 2-D gel is avoided, a greater number of proteins in the mixture can be represented. The disadvantage is that it can require an immense amount of time and computing power to deconvolute the data obtained. In addition, considerable time and effort may be expended in the analysis of uninteresting proteins. One of the most exciting techniques to emerge as an alternative to protein electrophoresis is that of isotope-coded affinity tags (ICAT). This method allows the quantitative protein profiling between different samples without the use of electrophoresis.

Acquisition of Protein Structure Information

Edman sequencing

One of the earliest methods used for protein identification was microsequencing by Edman chemistry to obtain N-terminal amino acid sequences. Little has changed in Edman chemistry since its introduction, but improvements in sequencing technology have increased the sensitivity and ease of Edman sequencing .The N-terminal sequencing of proteins was introduced by Edman in 1949. Today, Edman sequencing is most often used to identify proteins after they are transferred to membranes. The development of membranes compatible with sequencing chemicals allowed Edman sequencing to become a more applicable sequencing method for the identification of proteins separated by SDS-polyacrylamide gel electrophoresis.

In mixed-peptide sequencing, a protein is converted into peptides by cleavage with cyanogen bromide (CNBr) or skatole and the peptides are sequenced in an Edman sequencer simultaneously. Briefly, the process of mixed-peptide sequencing involves

separation of a complex protein mixture by polyacrylamide gel electrophoresis (1-D or 2-D) and then transfer of the proteins to an inert membrane by electroblotting. The proteins of interest are visualized on the membrane surface, excised, and fragmented chemically at methionine (by CNBr) or tryptophan (by skatole) into several large peptide fragments. On average, three to five peptide fragments are generated, consistent with the frequency of occurrence of methionine and tryptophan in most proteins. The membrane piece is placed directly into an automated Edman sequencer without further manipulation. Between 6 and 12 automated Edman cycles are carried out (4 to 8 h), and the mixed-sequence data are fed into the FASTF or TFASTF algorithms, which sort and match the data against protein (FASTF) and DNA (TFASTF) databases to unambiguously identify the protein. A recent variation of T/FASTF has been devised for MS. The T/FASTF/S programs are available at http://fasta.bioch.virginia.edu/.

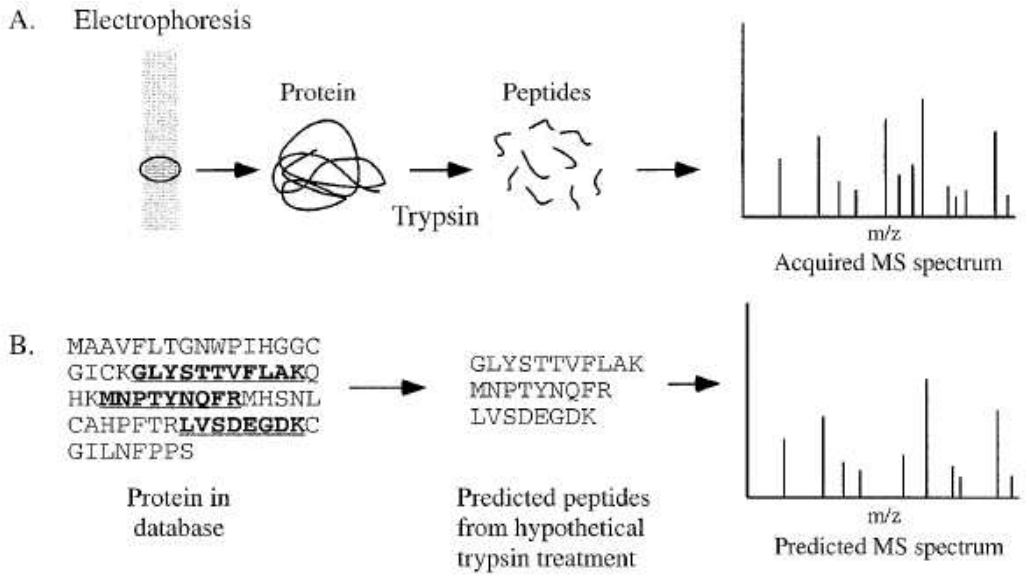


Fig: Strategy of protein identification by peptide mass fingerprinting. (A) The unknown protein is excised from a gel and converted topeptides by the action of a specific protease. The mass of the peptides produced is then measured in a mass spectrometer. (B) The mass spectrumof the unknown protein is searched against theoretical mass spectra produced by computer-generated cleavage of proteins in the database

Mass spectrometry

MS enables protein structural information, such as peptide masses or amino acid sequences, to be obtained. This information can be used to identify the protein by searching nucleotide and protein databases. It also can be used to determine the type and location of protein modifications. The harvesting of protein information by MS can be divided into three stages: (i) sample preparation, (ii) sample ionization, and (iii) mass analysis.

(i) Sample preparation. In most of proteomics, a protein is resolved from a mixture by using a 1- or 2-D polyacrylamide gel. The challenge is to extract the protein or its constituent peptides from the gel, purify the sample, and analyze it by MS. The extraction of whole proteins from gels is inefficient; however, if a protein is "in-gel" digested with a protease, many of the peptides can be extracted from the gel. A method for in-gel protein digestion was developed and is now commonly applied to both 1- and 2-D gels. In-gel digestion is more efficient at sample recovery than other common methods such as electroblotting. In addition, the conversions of a protein into its constituent peptides provide more information than can be obtained from the whole protein itself. For many applications, the peptides recovered following in-gel digestion need to be purified to remove gel contaminants. Common impurities from electrophoresis such as salts, buffers, and detergents can interfere with MS. In addition, peptide samples often require concentration before being analyzed by MS. One method of peptide purification commonly employed for this purpose is reverse-phase chromatography, which is available in a variety of formats. Peptides can be purified with ZipTips (Millipore) or Poros R2 perfusion material (PerSeptive Biosystems, Framingham, Mass.) or by high-pressure liquid chromatography (HPLC).

(ii) Sample ionization. For biological samples to be analyzed by MS, the molecules must be charged and dry. This is accomplished by converting them to desolvated ions. The two most common methods for this are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In both methods, peptides are converted to

ions by the addition or loss of one or more protons. ESI and MALDI are “soft” ionization methods that allow the formation of ions without significant loss of sample integrity. This is important because it enables accurate mass information to be obtained about proteins and peptides in their native states.

(iii) Mass analysis. Mass analysis follows the conversion of proteins or peptides to molecular ions. This is accomplished by the mass analyzers in a mass spectrometer, which resolve the molecular ions on the basis of their mass and charge in a vacuum.

(iv) Types of mass spectrometers.

Most mass spectrometers consist of four basic elements: (i) an ionization source, (ii) one or more mass analyzers, (iii) an ion mirror, and (iv) a detector. The names of the various instruments are derived from the name of their ionization source and the mass analyzer. Some of the most common mass spectrometers are discussed; for a more comprehensive review of mass spectrometers, the reader is directed to references. The analysis of proteins or peptides by MS can be divided into two general categories: (i) peptide mass analysis and (ii) amino acid sequencing.

In peptide mass analysis or peptide mass fingerprinting, the masses of individual peptides in a mixture are measured and used to create a mass spectrum. In amino acid sequencing, a procedure known as tandem mass spectrometry, or MS/MS, is used to fragment a specific peptide into smaller peptides, which can then be used to deduce the amino acid sequence.

MALDI-TOF. The principal application of a MALDITOF mass spectrometer is peptide mass fingerprinting because it can be completely automated, making it the method of choice for large-scale proteomics work. Because of its speed, MALDI-TOF is frequently used as a first-pass instrument for protein identification. If proteins cannot be identified by fingerprinting, they can then be analyzed by electrospray and MS/MS. A MALDI-TOF machine can also be used to obtain the amino acid sequence of peptides by a method known as post-source decay. However, peptide sequencing by post-source decay is not as reliable as sequencing with competing electrospray methods because the peptide fragmentation patterns are much less predictable. In MALDI, the sample is incorporated into matrix molecules and then subjected to irradiation by a laser. The laser

promotes the formation of molecular ions. The matrix is typically a small energy-absorbing molecule such as 2,5-dihydroxybenzoic acid or cyano-4-hydroxycinnamic acid. The analyte is spotted, along with the matrix, on a metal plate and allowed to evaporate, resulting in the formation of crystals. The plate, which an be 96-well format, is then placed in the mass spectrometer, and the laser is automatically targeted to specific places on the plate. Since sample application can be performed by a robot, the entire process including data collection and analysis can be automated. This is the single biggest advantage of MALDI. Another advantage of MALDI over ESI is that samples can often be used directly without any purification after in-gel digestion.

Conclusion and future prospects

Most of proteomics relies on methods, such as protein purification or PAGE, that are not high-throughput methods. Even performing MS can require considerable time in either data acquisition or analysis. Although hundreds of proteins can be analyzed quickly and in an automated fashion by a MALDI-TOF mass spectrometer, the quality of data is sacrificed and many proteins cannot be identified. Much higher quality data can be obtained for protein identification by MS/MS, but this method requires considerable time in data interpretation. In our opinion, new computer algorithms are needed to allow more accurate interpretation of mass spectra without operator intervention.New technologies will have to emerge before protein analysis on a large-scale (such as mapping the human proteome) becomes a reality.

Another major challenge for proteomics is the study of low abundance proteins. In some eukaryotic cells, the amounts of the most abundant proteins can be 106-fold greater than those of the low-abundance proteins. Many important classes of proteins (that may be important drug targets) such as transcription factors, protein kinases, and regulatory proteins are low-copy proteins. These low-copy proteins will not be observed in the analysis of crude cell lysates without some purification.

Reference

Paul, R. G. and Timothy, A. J. H. 2002. Molecular Biologist's Guide to Proteomics. Microbiology and molecular biology reviews 66 (1), 39-63.

Protein Sequence and Structure Database

Ms. N.Bharathi,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
bharathi_bioinfo@yahoo.co.in

Introduction

Bioinformatics is the application of Information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins (the building blocks of organisms) and nucleic acids (the information carrier). The biological information of nucleic acids is available as sequences while the data of proteins is available as sequences and structures. Sequences are represented in single dimension where as the structure contains the three dimensional data of sequences. A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.

The activity of preparing a database can be divided into:

- Collection of data in a form which can be easily accessed
- Making it available to a multi-user system (always available for the user)

Biological Databases

When Sanger first discovered the method to sequence proteins, there was a lot of excitement in the field of Molecular Biology. Initial interest in Bioinformatics was propelled by the necessity to create databases of biological sequences.

As of 2006, there are over 1,000 public and commercial biological databases. These biological databases usually contain genomics and proteomics data, but databases are also used in taxonomy. The data are nucleotide sequences of genes or amino acid sequences of proteins. Furthermore information about function, structure, localization on chromosome, clinical effects of mutations as well as similarities of biological sequences can be found.

Biological databases have become an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of

biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.Biological databases can be broadly classified in to sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only proteins.

Primary databases

Contain sequence data such as nucleic acid or protein**.** Example of primary databases includes:

1. PIR Protein Information Resource (Georgetown University Medical Center (GUMC))
2. Swiss-Prot Protein Knowledgebase (Swiss Institute of Bioinformatics)

Protein information resource (PIR)

The PIR is a division of the National Biomedical Research Foundation (NBRF) in the United States. It is involved in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japanese International Protein Sequence Database (JIPID) (15).

PIR grew out of Margaret Dayhoff's work in the middle of the 1960s. It strives to be comprehensive, well organized, accurate and consistently annotated. However, it is generally believed that it does not reach the level of completeness in the entry annotation, as does SWISS-PROT. Although SWISS-PROT and PIR overlap extensively, there are still many sequences, which can be found in only one of them.

Dayhoff and coworkers organized the proteins into families and super-families based on the degree of sequence similarity. Tables that reflected the frequency of changes observed in the sequences of a group of closely related proteins were then derived. Proteins that were less than 15% different were chosen to avoid the chance that the observed amino acid changes reflected two sequential amino acid changes instead of only

one. From aligned sequences, a phylogenetic tree was derived showing graphically which sequences were most related and therefore shared a common branch on the tree. Once the trees were made, they were used to score the amino acid changes that occurred during evolution of the genes for these proteins in the various organisms from which they originated.

Subsequently, a set of matrices (tables) the percent amino acid mutations by evolutionary selection of PAM tables which showed the probability that one amino acid changed into any other in these trees was constructed, thus showing which amino acid are most conserved at the corresponding position in two sequences. These tables are still used to measure similarity between two sequences and in database searches to find sequences that match a query sequence. The rule used is that the more identical and conserved amino acid that there are in two sequences, the more likely they are to have been derived from a common ancestor gene during evolution. If the sequences are very much alike, the proteins probably have the same biochemical function and 3-D structural folds. Thus, Dayhoff and colleagues contributed in several ways to modern biological sequence analysis by providing the first protein sequence database as well as PAM tables for performing protein sequence comparisons. One can search for entries or do sequence similarity searches at the PIR site. The database can also be downloaded as a set of flat files.

SWISS-PROT and TrEMBL

The SWISS-PROT protein sequence database was begun at the University of Geneva in 1986, and since 1987, it has been produced in a joint collaboration with the EBI (13). Detailed collaboration with world experts has enabled SWISS-PROT to have accurate and comprehensive high quality annotation. It is SWISS-PROT that has pioneered the notion of providing extensive cross-references between biological information sources that help to create a network of interacting databases. However, proteins and nucleotide coding regions are being deposited at a faster rate than ever the SWISS-PROT teams can handle. To this end, the EBI developed TrEMBL as a supplement to SWISS-PROT. TrEMBL is an automatic computer annotated database of

the translations of coding domains in EMBL that are not currently in SWISS-PROT. The protein sequence annotation is derived from annotations of the nucleotide sequence, analogies with already understood proteins, plus references to patterns and motif characteristics of particular protein functions.The databases can be accessed and searched through the SRS system at ExPASy, or one can download the entire database as one single flat file.

Secondary databases

The secondary databases are also known as pattern databases which contains results from the analysis of the sequences in the primary databases

Example of secondary databases includes:

1. Prosite
2. Prints
3. Pfam
4. Prodom

1. Prosite

PROSITE is a method of determining what is the function of uncharacterized proteins translated from genomic or cDNA sequences. It consists of a database of biologically significant sites and patterns formulated in such a way that with appropriate computational tools it can rapidly and reliably identify to which known family of protein (if any) the new sequence belongs.

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment, but it can be identified by the occurrence in its sequence of a particular cluster of residue types which is variously known as a pattern, motif, signature, or fingerprint. These motifs arise because of particular requirements on the structure of specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity. These requirements impose very tight constraints on the evolution of those limited (in size) but important portion(s) of a protein sequence. The use of protein

sequence patterns (or motifs) to determine the function(s) of proteins is becoming very rapidly one of the essential tools of sequence analysis.

2. Prints

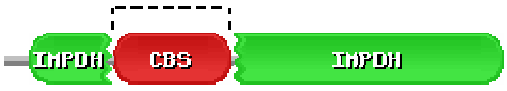
PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours.

3. Pfam

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.



74% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right. Pfam is a database of two parts, the first is the curated part of Pfam containing over 8296 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the

PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

4. Prodom

ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases

- ProDom *the whole database*
- ProDom-CG ***C****omplete* ***G****enomes only* (list)
- ProDom-SG for ***S****tructural* ***G****enomics*

Protein structure databases

1. Protein Data Bank (PDB) (Research Collaboratory for Structural Bioinformatics (RCSB))
2. CATH Protein Structure Classification
3. SCOP Structural Classification of Proteins
4. ModBase Database of Comparative Protein Structure Models

Protein Data Bank

The PDB is operated by the Research Collaboratory for Structural Bioinformatics (RCSB) under a contract to the U.S. National Science Foundation and is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students. The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data.Founded in 1971 by Brookhaven National Laboratory, management of the Protein Data Bank was transferred in 1998 to members of the Research Collaboratory for Structural Bioinformatics

(RCSB).The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

RCSB:

The **Research Collaboratory for Structural Bioinformatics (RCSB)** is a non-profit consortium dedicated to improving our understanding of the function of biological systems through the study of the 3-D structure of biological macromolecules. RCSB members work cooperatively and equally through joint grants and subsequently provide free public resources and publications to assist others and further the fields of bioinformatics and biology.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data. This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive

SCOP: (Structural Classification of Proteins)

In this database maintained at the MRC laboratory of molecular biology and centre for protein engineering describes structural and evolutionary relationships between proteins of known as structure (Murzin et al.,1995).because current automatic structure comparison tools cannot reliably identify all such relationships ,SCOP has been constructed using a combination of manual inspection and automated methods . The task is complicated by the fact that protein structures show such variety , ranging from small ,single domains to vast multi-domain assemblies . In some cases (e.g., some modular proteins),it may be meaningful to discuss a protein structure at the same time both at the multi-domain level and at the level of it's individual domains .

SCOP utilises a hierarchical scheme to organise the classification, allowing a four character code to be assigned to any protein domain.

The four classification levels are:

Class - A very broad description of the structural content of the protein

Fold - Indicative of a broad structural similarity but with no evidence of a homologous relationship

Super family - Sufficient structural similarity to infer a divergent evolutionary relationship but no detectable sequence similarity

Family - Significant sequence similarity which can be detected either directly or through a transitive search.

Domains - Autonomously-folding units of compact structure

SCOP defines the following categories, the first four of which are "true" structural

- proteins with only α-helices
- proteins with only β-sheets
- proteins with both α-helices and mainly parallel β-sheets (as beta-alpha-beta units)
- proteins with both α-helices and mainly antiparallel β-sheets (as separate alpha and beta domains)

Multidomain proteins

- membrane and cell surface proteins and peptides (not including those involved in the immune system)
- small proteins with well-defined structure
- coiled-coil proteins
- low-resolution protein structures
- peptides and fragments
- designed proteins of non-natural sequence

CATH:

The CATH database is a hierarchical domain classification of protein structures in the Protein Data Bank (PDB, Berman *et al.* 2003). Only crystal structures solved to resolution better than 4.0 angstroms are considered, together with NMR structures. All

non-proteins, models, and structures with greater than 30% "C-alpha only" are excluded from CATH. This filtering of the PDB is performed using the SIFT protocol (Michie *et al.,* 1996). Protein structures are classified using a combination of automated and manual procedures. There are four major levels in this hierarchy: **C**lass, **A**rchitecture, **T**opology (fold family) and **H**omologous superfamily (Orengo *et al.*, 1997). Each level is described below, together with the methods used for defining domain boundaries and assigning structures to a specific family.

The CATH Protein Structure Classification is a semi-automatic, hierarchical classification of protein domains published in 1997 by Christine Orengo, Janet Thornton and their colleagues. CATH shares many broad features with its principal rival, SCOP; however there are also many areas in which the detailed classification differs greatly. The name CATH is an acronym of the four main levels in the classification.

The four main levels of the CATH hierarchy are as follows:

Class, C-level:

Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognised; mainly-alpha, mainly-beta and alpha-beta. This last class (alpha-beta) includes both alternating alpha/beta structures and alpha+beta structures, as originally defined by Levitt and Chothia (1976). A fourth class is also identified which contains protein domains which have low secondary structure content.

Architecture, A-level*:*

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propellor or alpha four helix bundle).

Topology (Fold family), T-level*:*

Structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures. This is done using the structure

comparison algorithm SSAP (Taylor & Orengo, 1989) and CATHEDRAL (Harrison *et al.* 2002, 2003). Parameters for clustering domains into the same fold family have been determined by empirical trials throughout the databank (Orengo *et al.* 1992; Orengo *et al.* 1993; Harrison *et al.* 2002, 2003). Structures which have a SSAP score of 70 and where at least 60% of the larger protein matches the smaller protein are assigned to the same T level or fold group. Some fold fgroups are very highly populated (Orengo *et al.* 1994); Orengo & Thornton, 2005) particularly within the mainly-beta 2-layer sandwich architectures and the alpha-beta 3-layer sandwich architectures.

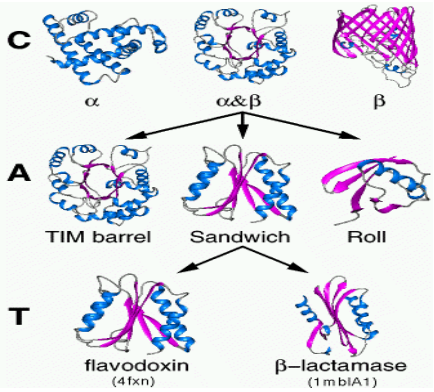
Homologous Superfamily, H-level*:*

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified either by high sequence identity or structure comparison using SSAP. Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria.

- Sequence identity > = 35%, overlap >= 60% of larger structure equivalent to smaller.
- SSAP score > = 80.0, sequence identity >= 20%, 60% of larger structure equivalent to smaller.
- SSAP score > = 70.0, 60% of larger structure equivalent to smaller, and domains which have related functions, which is informed by the literature and Pfam protein family database, (Bateman *et al.,* 2004).
- Significant similarity from HMM-sequence searches and HMM-HMM comparisons using SAM (Hughey &Krogh, 1996), HMMER (http://hmmer.wustl.edu) and PRC (http://supfam.org/PRC).
- **CATH** *defines* **four classes**: mostly-alpha, mostly-beta, alpha and beta, few secondary structures.
- In order to better understand the CATH classification system it is useful to know how it is constructed: much of the work is done by automatic methods, however there are important manual elements to the classification.

- The very first step is to separate the proteins into domains. It is difficult to produce an unequivocal definition of a domain and this is one area in which CATH and SCOP differ.
- The domains are automatically sorted into classes and clustered on the basis of sequence similarities. These groups form the H levels of the classification. The topology level is formed by structural comparisons of the homologous groups. Finally, the Architecture level is assigned manually.
- More detail on this process and the comparison between SCOP, CATH and FSSP can be found in: Hadley & Jones, 1999 (PMID 10508779) and Day et al., 2003 (PMID 14500873)

Pictorial represents of CATH:



Class:

- alpha domains only
- beta domains only
- alpha and beta

Architecture:

- roll
- TIM barrel
- sandwich

Topology:

- flavodoxin
- beta lactamase

References

Attwood. T. K, Parry-Smith. D. J and Phukan S. 2008. Introduction to Bioinformatics. Pearson Education Asia. New Delhi. 237.

Higgins. D and Taylor. W. 2000. Bioinformatics: Sequence, structure and databanks. Oxford University Press. Oxford. 249.

Mount. D. W. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press. New York. 564.

Online Resources

1. www.pir.georgetown.edu
2. www. expasy.org/sprot/
3. http://www.cathdb.info/latest/index.html
4. http://scop.mrc-lmb.cam.ac.uk/scop/
5. http://www.rcsb.org/pdb/

Applications of Expression Profiling Tools in Crop Improvement

Dr. M.Raveendran, Associate Professor,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
raveendrantnau@gmail.com

Genetic variation in the germplasm can be exploited by two ways to identify the genes/pathways governing the target traits. 1. QTL mapping of specific traits which involves the identification of putative chromosomal regions governing the traits 2) Taking advantage of our improved knowledge on anatomy and physiology, use of new molecular tools like proteomics and microarray to select particular tissues at specific times for analysis will reveal the differential expression of genes/proteins due to changes in the hormonal level or changes in physiological status. Comparison of expression profile between tolerant and sensitive genotypes will reveal the successful adaptations exhibited by the tolerant genotypes and failure of changes leading to susceptibility in the sensitive ones.

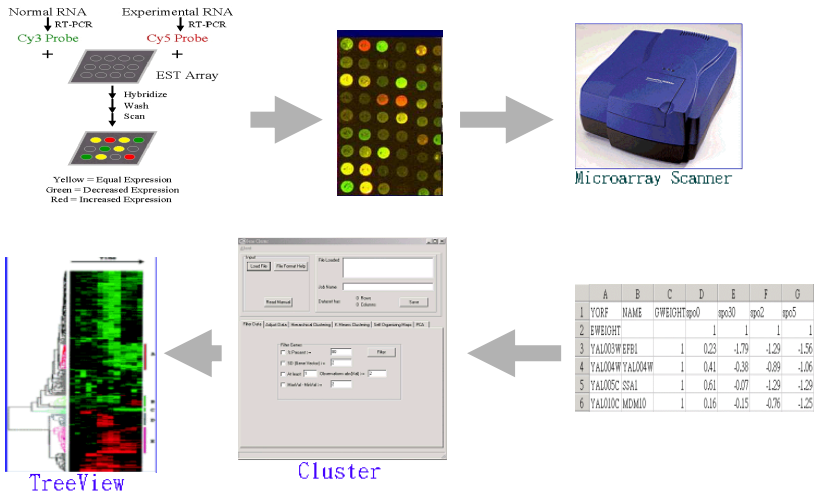
Previously techniques involving generation and characterization of Expressed Sequenced Tag's were used for understanding stress response. EST sequencing, as important as it continues to be for providing a fast overview of abundant transcripts in a species, is being superseded by more sophisticated approaches. Among these are libraries that are enriched for full-length cDNAs, and normalized and subtracted cDNA libraries. Such libraries can be made up of transcripts from multiple experimental conditions in which each RNA is converted into cDNA with an oligonucleotide tag attached that identifies a transcript population. Especially with fully sequenced genomes, SAGE provides, with little effort, hundreds of thousands of tags with each tag identifying the nature and abundance of a transcript. Conceivably, further advances in SAGE profiling could make this technique a viable alternative to other transcript profiling platforms. Transcript profiling, using GeneChips or long-oligonucleotide array slides, provides important insights into the dynamics of the transcriptional changes that accompany abiotic stress treatments. While both array platforms provide comparable results, they are distinguished by their ease of handling, dynamic range, reproducibility, and sensitivity to

nucleotide polymorphisms in the targets. Several contributions have reported on changes to transcript profiles under cold, drought, and high-salinity conditions, mainly in *A. thaliana* and rice and the variety of species in which analyses focusing on stress responses have been undertaken is gradually increasing. The precision of global expression profiling is now such that it can report with high confidence stress-dependent changes in transcript abundance that had previously been observed in isolated experiments. Due to the large number of genes involved in the response to various abiotic stresses, microarrays are increasingly being used to monitor global gene expression changes in Arabidopsis (Arabidopsis thaliana; Seki et al., 2002), rice (Kawasaki et al., 2001), and other systems (Ozturk et al., 2002).

Principle of expression microarrays:

The principle and terms used in microarray technology are the reverse of those used in traditional blot techniques. Thousands of gene specific DNA probe molecules are immobilized on the surface of a non-porous glass slide and hybridized against fluorescent labeled target molecules that represent mRNAs in the tissue being examined. The amount of fluorescent labels bound to each DNA probe on the array provides a direct estimate of gene expression in the tissue of interest.

Microarray processing sequence



A case study:

In a case study we compared drought responsive gene expression profiles in leaves of Apo, a moderately drought tolerant upland *indica* cultivar, and IR64, a drought susceptible lowland *indica* cultivar. These two genotypes are well known for their contrasting sensitivity to drought at vegetative stage under field conditions (Atlin et al., 2004). We reasoned that by sampling mRNA at critical physiological stage after exposure to water stress, we can identify expression patterns relevant to the phenotypic expression of tolerance to water stress. Expression profiling was done by using rice-22K Oligoarray from Agilent Technologies, USA. Comparisons of gene expression profiles between the two genotypes revealed genes and pathways that are consistent with observed physiological phenotypes before and after the onset of drought stress. Interestingly, we observed the patterns of correlated expression along the chromosomes and some of these regions were found co-localized with drought response QTL reported in the literature. Together, the gene expression data from these genotypes provide insights into physiological, biochemical and molecular changes that occur in rice leaves under drought stress and a source of candidate genes that are worthy of further investigation by genetic analysis.

Protein profiling

The importance of protein profiling has long been acknowledged in plant abiotic stress studies. Previous studies have provided useful information on individual enzymes or transporters, measuring their stress-dependent changes in quantity and activity. The results of such studies have formed the current hypotheses on stress responsive networks, in which protein modifications, protein–protein interactions, stress dependent protein movements, de novo synthesis, and controlled degradation play significant roles. The need for protein studies is also underscored, for example, by the recognition that rapid calcium spikes (which may be modulated in space, amplitude and frequency) lead to protein modifications that precede transcript changes. Consequently, large-scale high-throughput proteome analyses must be integrated with transcriptome and metabolome analyses if we are to obtain a comprehensive understanding of the stress response. In reality, large-scale proteome studies are still limited for several reasons. At present, there

are two major approaches to proteome profiling. The first and traditional approach uses two-dimensional gel separation (2D-PAGE) and mass spectrometry to measure changes in protein quantity (Yan et al., 2005). Although in excess of 1000 proteins can be readily identified, few stress-dependent changes in the quantities of these proteins have be identified, possibly due to problems in isolation or quantification or the inability to recognize protein modification. A second approach focuses on specific modifications of the proteome, such as membrane protein phosphorylation or populations of nitrosylated proteins. Unexpected, yet illuminating results have emerged from this approach, such as very divergent phosphorylation sites of receptor like- kinases from the same subfamily or the discovery of candidates for NO signaling pathways. Unlike transcriptome analyses, for which mature platforms exist, proteome analyses require novel, specific tool boxes. A specific protein chip has been developed for large-scale kinase assays to study the potential substrates of AtMPK3 and AtMPK6, the two mitogen associated protein kinases (MAPKs) known for their involvement in various stress responses The identified substrates contained a large number of ribosomal proteins. One interpretation and hypothesis would assume that, when under stress, the MAPKs might directly affect mRNA-loading by changing the phosphorylation state of ribosomal proteins.

A case study:

Proteomic approaches to understand drought responsiveness in peduncle tissues of rice

Rice's susceptibility to water stress is more pronounced at the reproductive stage and causes the maximum reduction in grain yield when stress coincides with the irreversible reproductive processes (Matsushima, 1966; Cruz and O'Toole, 1984). Possible reasons for this yield reduction includes poor panicle emergence due to incomplete peduncle elongation, pollination abnormalities and incomplete seed maturation. The genetic differences in peduncle behavior under drought stress and re-watering could be due to variations in the production, transport, breakdown, reception or signal transduction of hormones.

Plant material:

Rice plants (Variety: IR64) were grown under green house conditions. Drought stress was given for 3 days to a set of plants starting 3DBH. Observations on peduncle length, ABA content, final peduncle length and spikelet fertility were recorded. Peduncle tissues under control, drought and rewatered conditions were collected and used for proteomic studies.

Results:

The peduncle elongation was found to be very rapid (3-4 cm per day) under normal conditions and at the time of heading, the peduncle length was around 10 cm. When the plants were subjected to drought (for 3 days starting from 3 days before heading), the peduncle elongation was affected and delayed the heading process. Upon rewatering the peduncle tissues started re-elongation but the intermittent drought had significant effect on the final length of the peduncle (25 cm) when compared to the respective well-watered controls (25 cm; Ji et al., 2005, Table 1).

Proteomic analysis of peduncle tissues upon drought and rewatering

The two-dimensional gel electrophoresis of the peduncle tissue proteins collected under well watered, drought stressed and rewatered conditions revealed the reversible and irreversible behavior of the proteins during drought and rewatering.

Affected proteins are associated with growth processes like cell division (TCTP), cell elongation (Bet V allergen like protein), cell wall synthesis (XTH), ethylene and lignin biosynthesis (SAM), ABA sensitivity (LEA), cytoskeleton (ADF) etc., Six proteins are reported to be ABA responsive.

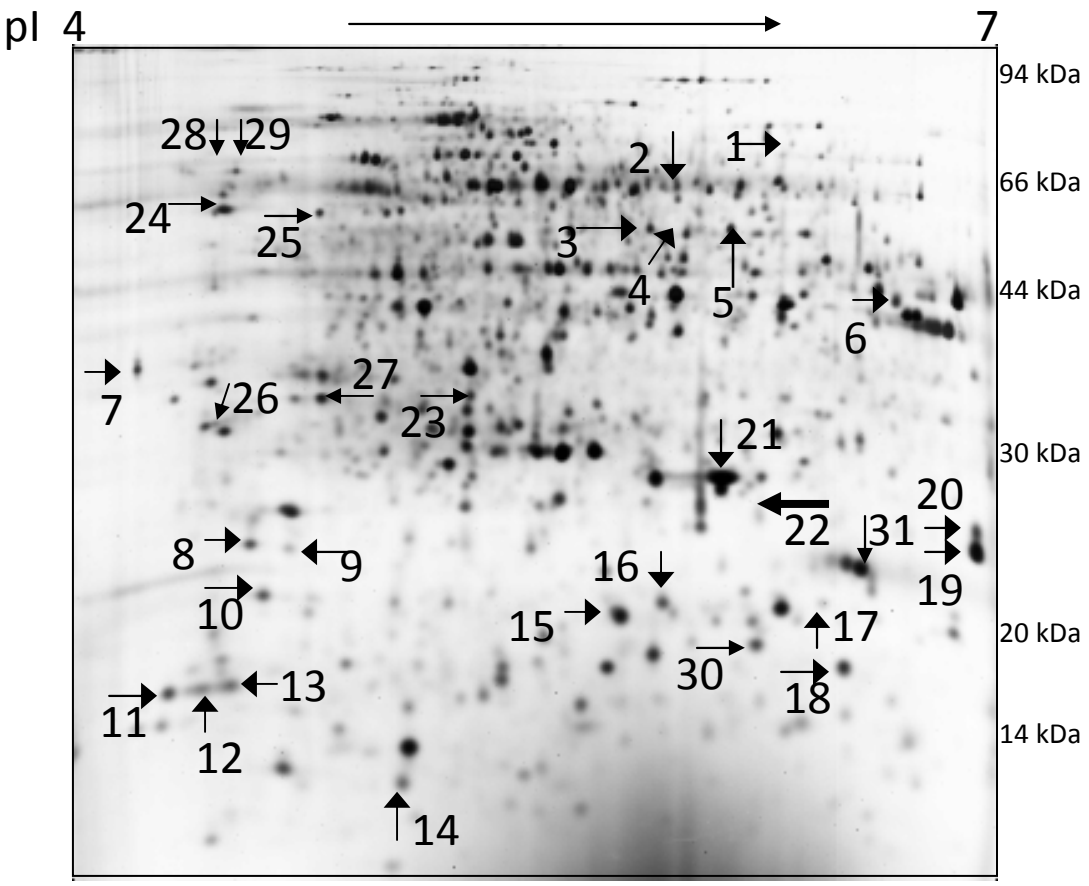


Fig. 2-D gel analysis of proteins extracted from the peduncle tissues of IR 64 harvested under drought (stressed for 3 days starting from 3 days before heading). In the first dimension (IEF), 100 µg of protein was loaded on an 18 cm IPG strip with a linear gradient of pH 4-7. In the second dimension, 12% SDS-PAGE gels were used with M_r standards. Proteins were visualized by silver staining. The arrows indicate 22 proteins that showed reversible and irreversible changes reproducibly and significantly under drought and rewatered conditions.

References:

Atlin et al., IRRI, Personal communication.

Ji, X.M., Raveendran, M., Oane, R., Ismail, A., Lafitte, R., Bruskiewich, R., Cheng, SH and J. Bennett. 2005. Tissue-specific expression and drought responsiveness of cell-wall invertase genes of rice at flowering. *Plant Molecular Biology***,** 59, 945-964.

Kawasaki S, Borchert C, Deyholos M, Wang H, Brazille S, Kawai K, Galbraith D, Bohnert HJ (2001) Gene expression profiles during the initial phase of salt stress in rice. *Plant Cell 13*, 889–905.

Liu.J., Raveendran.M**.,** Mushtaq.R., Xuemei.J., Xiaoe.Y., Bruskiewich.R., Katiyar.S., Shihua.C., Lafitte.R and Bennett.J. 2005. Proteomic analysis of drought-responsiveness in rice: OsADF5. In: *Green revolution to Gene revolution*, Bologna, Italy.

Ozturk ZN, Talame V, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N, Tuberosa R, Bohnert HJ (2002) Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant Mol Biol. 48*, 551–573.

Raveendran. M., Bruskiewich.R and Bennett.J. 2004. Proteomic analysis of drought responsiveness in rice. Rockefeller Foundation, USA sponsored Drought Conference held at CIMMYT Mexico. (May 23-27).

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T **(**2002). Monitoring the expression pattern of around 7000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. *Funct Integr Genomics. 2,*282-291.

Practical Sessions

Bioinformatics Tools and its Applications – I

Dr.J.Ramalingam, Associate Professor,
V.Srividhya, Senior Research Fellow,
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
jrjagdish@tnau.ac.in.

The growth of biotechnology has increased exponentially during the last decade as a result of development of high throughput techniques and parallel progress in information technology, which has led to the emergence of a novel discipline called bioinformatics. Biology is now very much an information science and bioinformatics provides the means to connect biological data to hypotheses. Bioinformatics tools and applications provide up-to-date descriptions of the various areas of applied bioinformatics, from the analysis of sequence, literature, and functional data to the function and evolution of organisms. The ability to process and interpret large volumes of data is essential with the application of new high throughput DNA sequencers providing an overload of sequence data. This provides an introduction to the analysis of DNA and protein sequences, from motif detection to gene prediction and annotation, with specific chapters on DNA and protein databases as well as data visualization.

Sequence retrieval

This exercise will focus on how to retrieve the sequences of both protein and nucleic acid from primary sequence databases like NCBI, EMBL and Swissprot

NCBI

National Center for Biotechnology Information was established in the year 1988. It provides databases such as Literature Databases, Entrez databases, Nucleotide Databases, Genome-Specific Resources and also tools for Data mining, Sequence analysis, 3-D Structure Display and Similarity searching, etc.. Main goal of NCBI is to analyze the sequence of gene and gene products and to gain better understanding of organization of genes and also to predict the structure of molecules analyzed.

EMBL:

The EMBL (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. It can be accessed and searched through the SRS system at EBI.

NCBI:

(I) **To retrieve the Nucleotide Sequence:**

1. Open the NCBI home page.
2. Select Nucleotide from Search option.
3. Enter the organism name in the search text box
4. Click Go to get the result
5. Change the format of the sequence in the display option to retrieve the sequence in required format.
6. Save the Sequence retrieved.

(II) **To retrieve the Protein Sequence**:

1. Open the NCBI home page.
2. Select Protein from Search option.
3. Enter the name of the protein in the search text box
4. Click Go to get the result
5. Change the format of the sequence in the display option to retrieve the sequence required format.
6. Save the Sequence retrieved.

EMBL:

1. Open EMBL web page using the URL www.ebi.ac.uk/embl/
2. Select Database in eb- eye search option.
3. Give any organism name
4. You will find all information that are provided by EMBL, as well as from its ftp sites.
5. Select Nucleotide Sequences
6. Click on the EMBL accession number in the resulting page.
7. Result will be the information on that particular nucleotide sequence save it EMBL format.

Swiss-Prot

Swiss-Prot is a manually curated biological database of protein sequences. Swiss-Prot was created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. Swiss-Prot strives to provide reliable protein sequences associated with a high level of annotation such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc. And also minimal level of redundancy and high level of integration with other databases. In 2002, the UniProt consortium was created: it is a collaboration between the Swiss Institute of Bioinformatics, the European Bioinfomatics Institute and the Protein Information Resource (PIR), funded by the National Institutes of Health. Swiss-Prot and its automatically curated supplement TrEMBL, have joined with the Protein Information Resource protein database to produce the UniProt Knowledgebase, the world's most comprehensive catalogue of information on proteins. As of 3 April 2007, UniProtKB/Swiss-Prot release 52.2 contains 263,525 entries.

1. Open the URL: www.expasy.org/sprot/
2. Search protein sequence by valid keyword such as entry name (ID), description (DE), gene name (GN), species (OS) or organelle (OG).
3. The results will be the sequences from UniProtKB/Swiss-Prot followed by UniProtKB/TrEMBL entries.

Compositional analysis of nucleotides

The nucleotides may be considered one of the most important metabolites of the cell. Nucleotides are found primarily as the monomeric units comprising the major nucleic acids of the cell, RNA and DNA.

Molecular weight depends on the number of nitrogen bases of nucleic acid (since the phosphates and sugars are a constant). Most DNA exists in the famous form of a double helix, in which two linear strands of DNA are wound around one another. The major force promoting formation of this helix is complementary base pairing: A's form hydrogen bonds with T's (or U's in RNA), and G's form hydrogen bonds with C's.

Reverse complement of a DNA/RNA sequence is that DNA/RNA sequence derived by reading the original sequence backwards and exchanging each nucelotide with that of its complement.

Codons are triplets of nucleotides that together specify an amino acid residue in a polypeptide chain. Most organisms use 20 or 21 amino acids to make their polypeptides, which are proteins or protein precursors. As there are four possible nucleotides, adenine (A), guanine (G), cytosine (C) and thymine (T) in DNA, there are 64 possible triplets to recognize only 20 amino acids plus the translation termination signal.

Codon usage is non random in a vast majority of prokaryotic and eukaryotic species. The major factor in codon choice in many unicellular and some multicellular organisms is Darwinian selection between synonyms; highly expressed genes using a restricted set of codons. This selection is almost certainly for optimal translational efficiency, and is most pronounced in highly expressed genes in species whose effective population size is large. Divergence of codon usage and choice of optimal codons correlates with evolutionary distance, but usage patterns in phylogenetically distant species may converge due to the similarities of factors that influence the drift in choice of optimal codons. Analysis of codon usage has been used to identify highly expressed genes. Codon bias is the probability that a given codon will be used for an amino acid over a different codon which codes the same amino acids.

Molecular weight and Base statistics:

This program will take a nucleic acid sequence and calculate the molecular weight and GC content. The program will ignore numbers, spaces or characters like B or Z which do not correspond to one of the 5 DNA bases. The atomic weights for each atom used are from the International Union of Pure and Applied Chemistry (IUPAC). The program assumes that each phosphate group has two hydrogen atoms bound to it, which may not be true outside the physiological pH range. This version can also deal with FASTA format sequences. It ignores any line of text which is started by a ">" character.

1. Open web page use of following URL http://www.encorbio.com/protocols/Nuc-MW.htm.
2. Input a nucleic acid sequence and calculate the molecular weight and base statistics.

Codon usage:

1. Open following URL http://www.bioinformatics.org/SMS/.
2. Go to codon usage from DNA analysis menu.
3. Input a nucleic acid sequence in FASTA format.

Complementary and reverse complimentary:

1. Open following URL http://www.bioinformatics.org/SMS/.
2. Go to reverse compliment from DNA manipulation menu.
3. For getting complimentary sequence select complimentary in drop down menu or select reverse complimentary for reverse complimentary sequence.
4. Input a nucleic acid sequence in FASTA format.

Compositional analysis of proteins

Molecular weight, the sum of the atomic weights of all atoms making up a molecule. Actually, what is meant by molecular weight is molecular mass. The use of this expression is historical, however, and will be maintained. The atomic weight is the mass, in atomic mass units, of an atom. It is approximately equal to the total number of nucleons, protons and neutrons composing the nucleus.

Isoelectric point (pI) is the pH at which a particular molecule or surface carries no net electrical charge. Amphoteric molecules called zwitterions contain both positive and negative charges depending on the functional groups present in the molecule. They are affected by pH of their surrounding environment and can become more positively or negatively charged due to the loss or gain of protons (H+). A molecule's pI can affect its solubility at a certain pH. Such molecules have minimum solubility at the pH which corresponds to their pI and are often seen to precipitate out of solution.

Biological amphoteric molecules such as proteins contain both acidic and basic functional groups. Amino acids which make up proteins may be positive, negative, neutral or polar in nature, and together give a protein its overall charge. At a pH below their pI, proteins carry a net positive charge. Above their pI they carry a net negative charge. Proteins can be separated according to their isoelectric point (overall charge) on a polyacrylamide gel using a technique called isoelectric focusing. This technique utilizes a pH gradient to separate proteins. Isoelectric focusing is also the

first step in performing 2-D gel polyacrylamide gel electrophoresis. The theoretical isoelectric point (pI) as a numerical vector of length one.

Instability index provides an estimate of the stability of your protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.

Aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. The aliphatic index of a protein is calculated according to the following formula:

Aliphatic index = X(Ala) + a * X(Val) + b * (X(Ile) + X(Leu))

Where X (Ala), X(Val), X(Ile), and X(Leu) are mole percent (100 X mole fraction) of alanine, valine, isoleucine, and leucine. The coefficients **a** and **b** are the relative volume of valine side chain (a = 2.9) and of Leu/Ile side chains (b = 3.9) to the side chain of alanine.

Hydropathy index of a protein is a number representing its hydrophilic or hydrophobic properties. The larger the number is, the more hydrophobic the amino acid. The most hydrophobic amino acids are isoleucine (4.5) and valine (4.2). The most hydrophilic ones are arginine (-4.5) and lysine (-3.9). This is very important in protein structure; hydrophobic amino acids tend to be internal (with regard to the protein's 3 dimensional shape) while hydrophilic amino acids are more commonly found towards the protein surface The Grand Average of Hydropathy value (GRAVY) for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence.

1. Open protparam webpage to calculate Mw, pl and residues percent by using the following URL. http://expasy.org/tools/protparam.html.
2. Give a protein sequence as input.

Sequence similarity

BLAST – Similarity search tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. The BLAST performs "local" alignments. The BLAST algorithm was developed by Altschul, Gish, Miller, Myers and Lipman in 1990**.** BLAST concentrates on finding regions of high local similarity in alignments without gaps, evaluated by an alphabet-weight scoring matrix.

1. Open the BLAST home Page using the URL
   http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
2. Use any of the following Basic BLAST program according to query:
   Nucleotide blast - Search a **nucleotide** database using a **nucleotide** query
   protein blast - Search **protein** database using a **protein** query
   blastx - Search **protein** database using a **translated nucleotide** query
   tblastn - Search **translated nucleotide** database using a **protein** query
   tblastx - Search **translated nucleotide** database using a **translated nucleotide** query
3. Enter the query sequence in the text box provided.
4. Click BLAST to get the result.

Bioinformatics Tools and its Applications- II

Gene finding in eukaryotes

The gene structure and the gene expression mechanism in eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of exons and introns. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called splicing takes place, in which the intron sequences are excised and discarded from the RNA sequence. The remaining RNA segments, the ones corresponding to the exons, are ligated to form the mature RNA strand.

Gene finding refers to the area of computational biology that is concerned with algorithmically identifying stretches of sequence, usually genomic DNA. This especially includes protein-coding genes, but may also include other functional elements such as RNA genes and regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced. Determining that a sequence is functional should be distinguished from determining the function of the gene or its product. The latter still demands in vivo experimentation through gene knockout and other assays, although frontiers of bioinformatics research are making it increasingly possible to predict the function of a gene based on its sequence alone.

Three types of information are used in predicting gene structures. one is "signals" in the sequence, such as splice sites; "content" statistics, such as codon bias; and similarity to known genes. The first two types have been used since the early days of gene prediction, whereas similarity information has been used routinely only in recent years. One of the reasons that the accuracy of gene-prediction programs have improved in the last few years is the enormous increase in the number of examples of known coding sequences.

1. Open following URL http://exon.gatech.edu/GeneMark/eukhmm.cgi
2. Email address is required for graphical output or sequences longer than 400000 bp in genemark tool.

Identifying ORF's in prokaryotes

An open reading frame or ORF is a portion of an organism's genome which contains a sequence of bases that could potentially encode a protein. Gene finding in prokaryotes is restricted to search for ORF. An ORF is a sequence of DNA that starts with start codon "ATG" and ends with any of the three termination codons (TAA, TAG, and TGA). Depending on the starting point, there are six possible ways (three on forward strand and three on complementary strand) of translating any nucleotide sequence into amino acid sequence according to the genetic code. These are called reading frames. ORFs are usually encountered when sifting through pieces of DNA while trying to locate a gene. Since there exist variations in the start-code sequence of organisms with altered genetic code, the ORF will be identified differently. A typical ORF finder will employ algorithms based on existing genetic codes and all possible reading frames.

1. Go to following URL www.ncbi.nlm.nih.gov/gorf/gorf.html
2. Nucleotide sequence should be FASTA format.

Basics of Linkage Map Construction

N. Manikanda Boopathi, Assistant Professor
Dept. of PMB & Biotechnology, CPMB, TNAU, Coimbatore 3.
nmboopathi@tnau.ac.in; biotechboopathi@yahoo.com

Generally genome mapping methods are divided into two categories: Genetic mapping and Physical mapping. However there is yet another map referred as - Cytogenetic map (a genetic term used to describe the visual appearance of a chromosome when stained and examined under a microscope).

Genetic or Linkage Mapping

As with any type of map, a genetic map must show the positions of distinctive features. In a geographic map these markers are recognizable components of the landscape, such as rivers, roads and buildings. What markers can we use in a genetic landscape?

Genes were the first markers to be used

The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly, used genes as markers. This was many years before it was understood that genes are segments of DNA molecules. Instead, genes were looked upon as abstract entities responsible for the transmission of heritable characteristics from parent to offspring. To be useful in genetic analysis, a heritable characteristic has to exist in at least two alternative forms or phenotypes, an example being tall or short stems in the pea plants originally studied by Mendel. Each phenotype is specified by a different allele of the corresponding gene. To begin with, the only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination. So, for example, the first fruit-fly maps showed the positions of genes for body color, eye color, wing shape and suchlike, all of these phenotypes being visible simply by looking at the flies with a low-power microscope or the naked eye. This approach was fine in the early days but geneticists soon realized that there were only a limited number of visual phenotypes whose inheritance could be studied, and in many cases their analysis was complicated because a single phenotype could be affected by more than one gene. To make gene maps more

comprehensive it would be necessary to find characteristics that were more distinctive and less complex than visual ones.

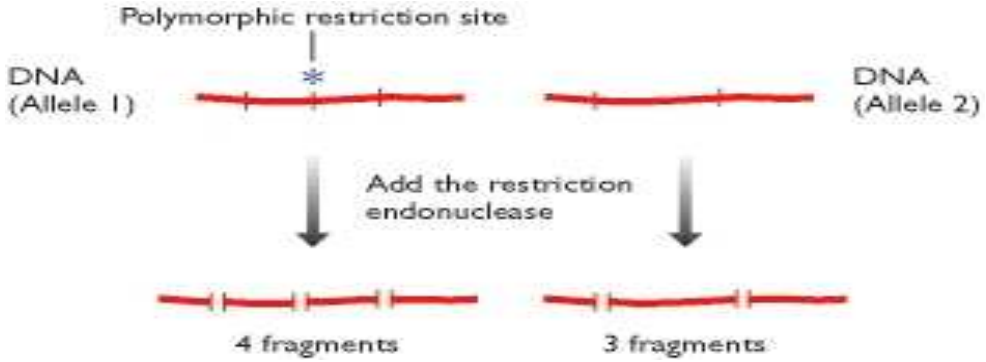
Biochemical markers

The answer was to use biochemistry to distinguish phenotypes. This has been particularly important with two types of organisms - microbes and humans. Microbes, such as bacteria and yeast, have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes.

DNA markers for genetic mapping

Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed. This would be true even if every gene could be mapped because, in most eukaryotic genomes the genes are widely spaced out with large gaps between them. The problem is made worse by the fact that only a fraction of the total number of genes exists in allelic forms that can be distinguished conveniently. Gene maps are therefore not very comprehensive. We need other types of marker.

Mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are several types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs) are few examples. Let us examine the principle of genetic map with the aid of Restriction fragment length polymorphisms (RFLPs)



Principle of restriction fragment length polymorphism (RFLP)

The DNA molecule on the left has a polymorphic restriction site (marked with the asterisk) that is not present in the molecule on the right. The RFLP is revealed after treatment with the restriction enzyme because one of the molecules is cut into four fragments whereas the other is cut into three fragments. RFLPs were the first type of DNA marker to be studied. Recall that restriction enzymes cut DNA molecules at specific recognition sequences. This sequence specificity means that treatment of a DNA molecule with a restriction enzyme should always produce the same set of fragments. This is not always the case with genomic DNA molecules because some restriction sites are polymorphic, existing as two alleles, one allele displaying the correct sequence for the restriction site and therefore being cut when the DNA is treated with the enzyme, and the second allele having a sequence alteration so the restriction site is no longer recognized. The result of the sequence alteration is that the two adjacent restriction fragments remain linked together after treatment with the enzyme, leading to a length polymorphism. This is an RFLP and its position on a genome map can be worked out by following the inheritance of its alleles, just as is done when genes are used as markers. There are thought to be about 10^5 RFLPs in the human genome, but of course for each RFLP there can only be two alleles (with and without the site). The value of RFLPs in human gene mapping is therefore limited by the high possibility that the RFLP being studied shows no variability among the members of an interesting family.

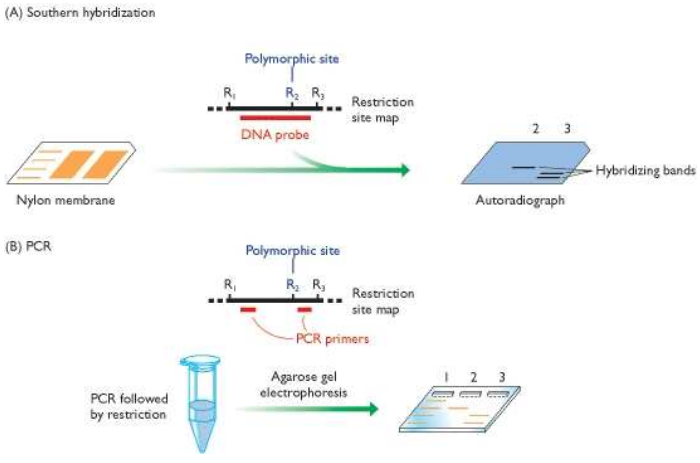


Figure. Methods for scoring an RFLP

(A) RFLPs can be scored by Southern hybridization. The DNA is digested with the appropriate restriction enzyme and separated in an agarose gel. The smear of restriction fragments is transferred to a nylon membrane and probed with a piece of DNA that spans the polymorphic restriction site. If the site is absent then a single restriction fragment is detected (lane 2); if the site is present then two fragments are detected (lane 3).

(B) The RFLP can also be typed by PCR, using primers that anneal either side of the polymorphic restriction site. After the PCR, the products are treated with the appropriate restriction enzyme and then analyzed by agarose gel electrophoresis. If the site is absent then one band is seen on the agarose gel; if the site is present then two bands are seen.

In order to score an RFLP, it is necessary to determine the size of just one or two individual restriction fragments against a background of many irrelevant fragments. This is not a trivial problem: an enzyme such as *Eco*RI, with a 6-bp recognition sequence, should cut approximately once every $4^6 = 4096$ bp and so would give almost 800 000 fragments when used with human DNA. After separation by agarose gel electrophoresis, these 800 000 fragments produce a smear and the RFLP cannot be distinguished. Southern hybridization, using a probe that spans the polymorphic restriction site, provides one way of visualizing the RFLP, but nowadays PCR is more frequently used. The primers for the PCR are designed so that they anneal either side of the polymorphic site, and the RFLP is typed by treating the amplified fragment with the restriction enzyme and then running a sample in an agarose gel.

Linkage analysis is the basis of genetic mapping

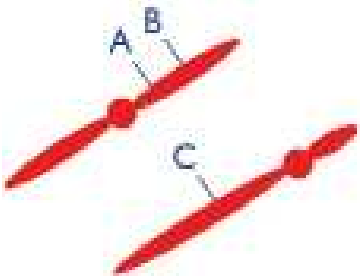
Now that we have assembled a set of markers with which to construct a genetic map we can move on to look at the mapping techniques themselves. These techniques are all based on genetic linkage, which in turn derives from the seminal discoveries in genetics made in the mid 19th century by Gregor Mendel.

The principles of inheritance and the discovery of linkage

Mendel studied seven pairs of contrasting characteristics in his pea plants, one of which was violet and white flower color. Genetic mapping is based on the principles of inheritance as first described by Gregor Mendel in 1865. From the results of his breeding experiments with peas, Mendel concluded that each pea plant possesses two alleles for each gene, but displays only one

phenotype. This is easy to understand if the plant is pure-breeding, or homozygous, for a particular characteristic, as it then possesses two identical alleles and displays the appropriate phenotype. However, Mendel showed that if two pure-breeding plants with different phenotypes are crossed then all the progeny (the F_1 generation) display the same phenotype. These F_1 plants must be heterozygous, meaning that they possess two different alleles, one for each phenotype, one allele inherited from the mother and one from the father. Mendel postulated that in this heterozygous condition one allele overrides the effects of the other allele; he therefore described the phenotype expressed in the F_1 plants as being dominant over the second, recessive phenotype. This is the perfectly correct interpretation of the interaction between the pairs of alleles studied by Mendel, but we now appreciate that this simple dominant-recessive rule can be complicated by situations that he did not encounter. One of these is incomplete dominance, where the heterozygous phenotype is intermediate between the two homozygous forms. An example is when red carnations are crossed with white ones, the F_1 heterozygotes being pink. Another complication is codominance, when both alleles are detectable in the heterozygote. Codominance is the typical situation for DNA markers.

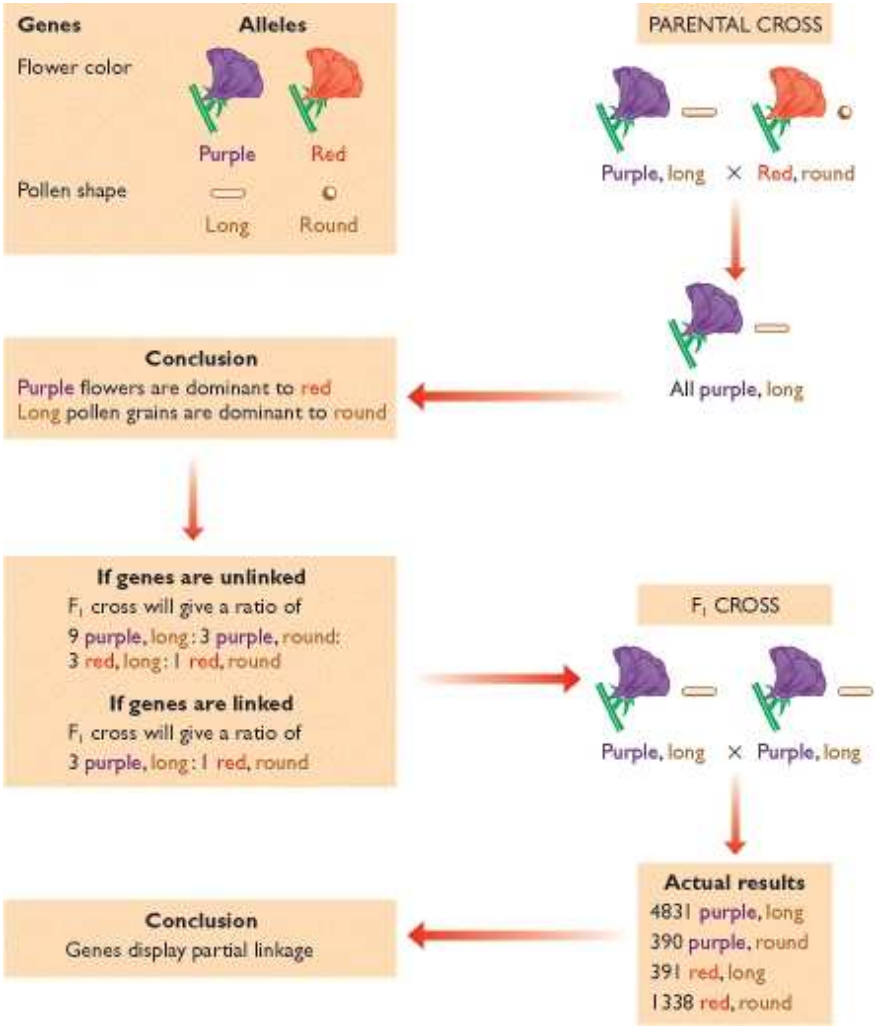
As well as discovering dominance and recessiveness, Mendel carried out additional crosses that enabled him to establish two Laws of Genetics. The First Law states that ***alleles segregate randomly***. In other words, if the parent's alleles are *A* and *a*, then a member of the F_1 generation has the same chance of inheriting *A* as it has of inheriting *a*). The Second Law is that ***pairs of alleles segregate independently***, so that inheritance of the alleles of gene A is independent of inheritance of the alleles of gene B. Because of these laws, the outcomes of genetic crosses are predictable.



When Mendel's work was rediscovered in 1900, his Second Law worried the early geneticists because it was soon established that genes reside on chromosomes, and it was realized that all organisms have many more genes than chromosomes. Chromosomes are inherited as intact units, so it was reasoned that the alleles of some pairs of genes will be inherited together because they are on the same chromosome as shown in the figure.

This is the principle of genetic linkage, and it was quickly shown to be correct, although the results did not turn out exactly as expected. The complete linkage that had been anticipated between many pairs of genes failed to materialize. Pairs of genes were either inherited

independently, as expected for genes in different chromosomes, or, if they showed linkage, then it was only partial linkage: sometimes they were inherited together and sometimes they were not. The resolution of this contradiction between theory and observation was the critical step in the development of genetic mapping techniques.



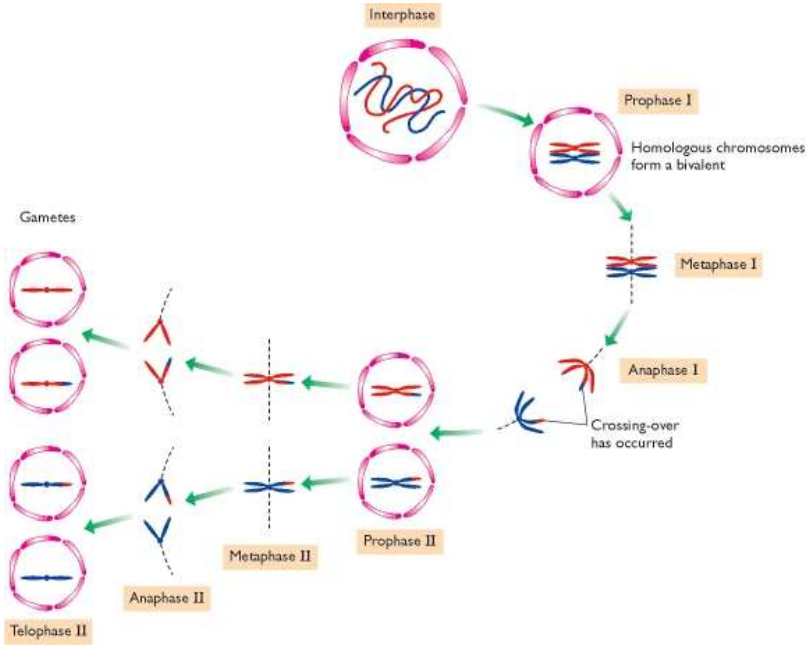
Partial linkage

Partial linkage was discovered in the early 20th century. The cross shown here was carried out by Bateson, Saunders and Punnett in 1905 with sweet peas. The parental cross gives the typical dihybrid result, with all the F_1 plants displaying the same phenotype, indicating that the dominant alleles are purple flowers and long pollen grains. The F_1 cross gives unexpected results as the progeny show neither a 9 : 3 : 3 : 1 ratio (expected for genes on different chromosomes) nor

a 3 : 1 ratio (expected if the genes are completely linked). An unusual ratio is typical of partial linkage.

Partial linkage is explained by the behavior of chromosomes during meiosis

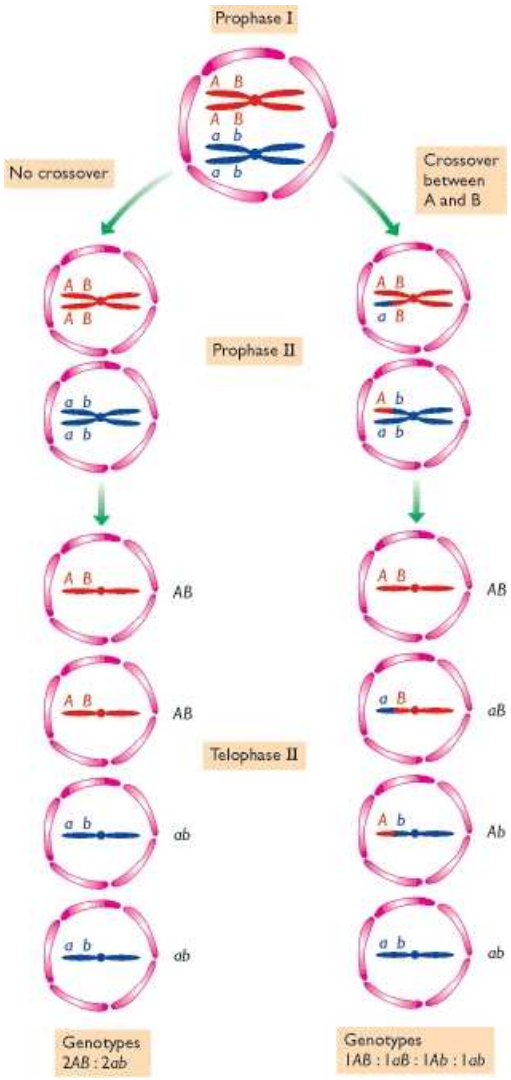
During interphase (the period between nuclear divisions) the chromosomes are in their extended form. At the start of mitosis the chromosomes condense and by late prophase have formed structures that are visible with the light microscope. Each chromosome has already undergone DNA replication but the two daughter chromosomes are held together by the centromere. During metaphase the nuclear membrane breaks down (in most eukaryotes) and the chromosomes line up in the center of the cell. Microtubules now draw the daughter chromosomes towards either end of the cell. In telophase, nuclear membranes re-form around each collection of daughter chromosomes. The result is that the parent nucleus has given rise to two identical daughter nuclei. For simplicity, just one pair of homologous chromosomes is shown; one member of the pair is red, the other is blue.



The critical breakthrough was achieved by Thomas Hunt Morgan, who made the conceptual leap between partial linkage and the behavior of chromosomes when the nucleus of a cell divides. Cytologists in the late 19th century had distinguished two types of nuclear division: mitosis and meiosis. Mitosis is more common, being the process by which the diploid nucleus of a somatic cell divides to produce two daughter nuclei, both of which are diploid. Before mitosis

begins, each chromosome in the nucleus is replicated, but the resulting daughter chromosomes do not immediately break away from one another. To begin with they remain attached at their centromeres and by cohesin proteins which act as 'molecular glue' holding together the arms of the replicated chromosomes. The daughters do not separate until later in mitosis when the chromosomes are distributed between the two new nuclei. Obviously it is important that each of the new nuclei receives a complete set of chromosomes, and most of the intricacies of mitosis appear to be devoted to achieving this end.

The events involving one pair of homologous chromosomes are shown; one member of the pair is red, the other is blue. At the start of meiosis the chromosomes condense and each homologous pair lines up to form a bivalent. Within the bivalent, crossing-over might occur, involving breakage of chromosome arms and exchange of DNA. Meiosis then proceeds by a pair of mitotic nuclear divisions that result initially in two nuclei, each with two copies of each chromosome still attached at their centromeres, and finally in four nuclei, each with a single copy of each chromosome. These final products of meiosis, the gametes, are therefore haploid.



Mitosis illustrates the basic events occurring during nuclear division but is not directly relevant to genetic mapping. Instead, it is the distinctive features of meiosis that interest us. Meiosis occurs only in reproductive cells, and results in a diploid cell giving rise to four haploid gametes, each of which can subsequently fuse with a gamete of the opposite sex during sexual reproduction. The fact that meiosis results in four haploid cells whereas mitosis gives rise to two diploid cells is easy to explain: meiosis involves two nuclear divisions, one after the other, whereas mitosis is just a single nuclear division. This is an important distinction, but the critical difference between mitosis and meiosis is more subtle. Recall that in a diploid cell there are two separate copies of each chromosome. We refer to these as pairs of homologous chromosomes. During mitosis, homologous chromosomes remain separate

from one another, each member of the pair replicating and being passed to a daughter nucleus independently of its homolog. In meiosis, however, the pairs of homologous chromosomes are by no means independent. During meiosis I, each chromosome lines up with its homolog to form a bivalent (see Figure). This occurs after each chromosome has replicated, but before the replicated structures split, so the bivalent in fact contains four chromosome copies, each of which is destined to find its way into one of the four gametes that will be produced at the end of the meiosis. Within the bivalent, the chromosome arms (the chromatids) can undergo physical breakage and exchange of segments of DNA. The process is called crossing-over or recombination and was discovered by the Belgian cytologist Janssens in 1909. This was just 2 years before Morgan started to think about partial linkage.

The effect of a crossover on linked genes

The drawing shows a pair of homologous chromosomes, one red and the other blue. A and B are linked genes with alleles *A*, *a*, *B* and *b*. On the left is a meiosis with no crossover between A and B: two of the resulting gametes have the genotype *AB* and the other two are *ab*. On the right, a crossover occurs between A and B: the four gametes display all of the possible genotypes: *AB*, *aB*, *Ab* and *ab*.

How did the discovery of crossing-over help Morgan explain partial linkage? To understand this we need to think about the effect that crossing-over can have on the inheritance of genes. Let us consider two genes, each of which has two alleles. We will call the first gene A and its alleles *A* and *a*, and the second gene B with alleles *B* and *b*. Imagine that the two genes are located on chromosome number 2 of *Drosophila melanogaster*, the species of fruit fly studied by Morgan. We are going to follow the meiosis of a diploid nucleus in which one copy of chromosome 2 has alleles *A* and *B*, and the second has *a* and *b*. This situation is illustrated in above Figure. Consider the two alternative scenarios:

1. ***A crossover does not occur between genes A and B.*** If this is what happens then two of the resulting gametes will contain chromosome copies with alleles *A* and *B*, and the other two will contain *a* and *b*. In other words, two of the gametes have the genotype *AB* and two have the genotype *ab*.

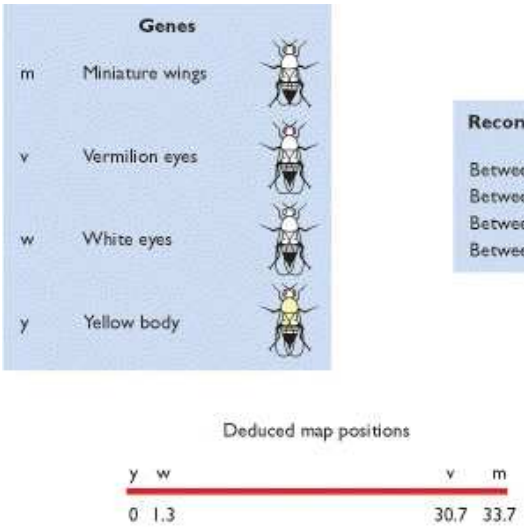
2. ***A crossover does occur between genes A and B.*** This leads to segments of DNA containing gene B being exchanged between homologous chromosomes. The eventual result is that each gamete has a different genotype: 1 *AB*, 1 *aB*, 1 *Ab*, 1 *ab*.

Now think about what would happen if we looked at the results of meiosis in a hundred identical cells. If crossovers never occur then the resulting gametes will have the following genotypes:

200 *AB*
200 *ab*

This is complete linkage: genes A and B behave as a single unit during meiosis. But if (as is more likely) crossovers occur between A and B in some of the nuclei, then the allele pairs will not be inherited as single units. Let us say that crossovers occur during 40 of the 100 meioses. The following gametes will result:

160 *AB*
160 *ab*
40 *Ab*
40 *aB*

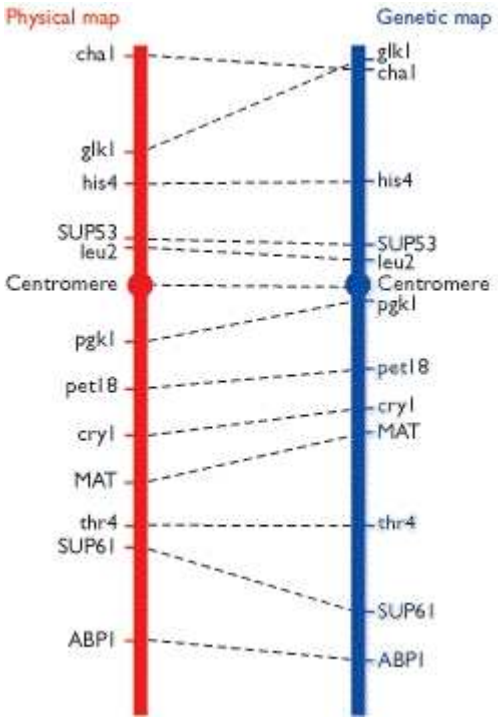


The linkage is not complete, it is only partial. As well as the two **parental** genotypes (*AB*, *ab*) we see gametes with recombinant genotypes (*Ab*, *aB*). From partial linkage to genetic mapping: Working out a genetic map from recombination frequencies

The example is taken from the original experiments carried out with fruit flies by Arthur Sturtevant. All four genes are on the X chromosome of the fruit fly. Recombination frequencies between the genes are shown, along with their deduced map positions.

Once Morgan had understood how partial linkage could be explained by crossing-over during meiosis he was able to devise a way of mapping the relative positions of genes on a chromosome. In fact the most important work was done not by Morgan himself, but by an undergraduate in his laboratory, Arthur Sturtevant (Sturtevant, 1913). Sturtevant assumed that

crossing-over was a random event, there being an equal chance of it occurring at any position along a pair of lined-up chromatids. If this assumption is correct then two genes that are close together will be separated by crossovers less frequently than two genes that are more distant from one another. Furthermore, the frequency with which the genes are unlinked by crossovers will be directly proportional to how far apart they are on their chromosome. The recombination frequency is therefore a measure of the distance between two genes. If you work out the recombination frequencies for different pairs of genes, you can construct a map of their relative positions on the chromosome.



Comparison between the genetic and physical maps of *Saccharomyces cerevisiae* chromosome III

The comparison shows the discrepancies between the genetic and physical maps, the latter determined by DNA sequencing. Note that the order of the upper two markers (glk1 and cha1) is incorrect on the genetic map, and that there are also differences in the relative positioning of other pairs of markers.

It turns out that Sturtevant's assumption about the randomness of crossovers was not entirely justified. Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by physical mapping and DNA sequencing, have shown that some regions of chromosomes, called recombination hotspots, are more likely to be involved in crossovers than others. This means that a genetic map distance does not necessarily indicate the physical distance between two markers. Also, we now realize that a single chromatid can participate in more than one crossover at the same time, but that there are limitations on how close together these crossovers can be, leading to more inaccuracies in the mapping procedure. Despite these qualifications, linkage analysis usually makes correct deductions about gene order, and distance estimates are sufficiently accurate to generate genetic maps that are of value as frameworks for genome sequencing projects.

References

- Bovenhuis, H. and T.H.E. Meuwissen. 1996. Detection and mapping of quantitative trait loci. Animal Genetics and Breeding Unit. UNE, Armidale, Australia. ISBN 186389 323 7
- Bulmer, M.G. 1971. The effect of selection on genetic variability. *Amer. Nat*. 105:201.
- Morton, N.E. 1955. Sequential tests for the detection of linkage. *American Journal of Human Genetics.* 7:277-318.
- http://www.scribd.com/doc/6229849/Mapping-Population
- http://www.generationcp.org/mab/index.php?id=134
- http://www.ncbi.org/help/linkageprinciples

Structured Association Mapping using STRUCTURE and TASSEL

Dr. K K Vinod,
IARI, Rice Breeding and Genetics Research Centre,
Aduthurai 612101, Tanjavur District.
kkvinodh@gmail.com

With advances in genotyping technology, including rapid increases in the number of genetic markers available for QTL studies, association analysis is now a viable approach for the dissection of complex genetic traits (Churchill et al. 2004). Association mapping involves assessment of population structure and using this population information and kinship information among individuals to assess marker – trait association. Two common software packages widely used today for association mapping are STRUCTURE (Pritchard et al. 2010) and TASSEL (Buckler et al. 2009). STRUCTURE implements a model-based clustering method for inferring population structure using genotype data consisting of unlinked markers. This program can demonstrate the presence of population structure, identify distinct genetic populations, assign individuals to populations, and identify migrants and admixed individuals. Trait Analysis by Association, Evolution and Linkage, or TASSEL, makes use of the most advanced statistical methods to maximize statistical power for finding QTL. Both a structured association approach (Pritchard *et al.* 2000; Thornsberry *et al.* 2001) and a unified mixed model method have been implemented to minimize the risk of false positives by integrating population structure and family relatedness within populations (Yu *et al.* 2006).

I. Determining population structure using Structure 3.2.2

A. Preparation of marker genotype data

Prepare a matrix of marker genotype data in Excel as given below, for microsatellite data:

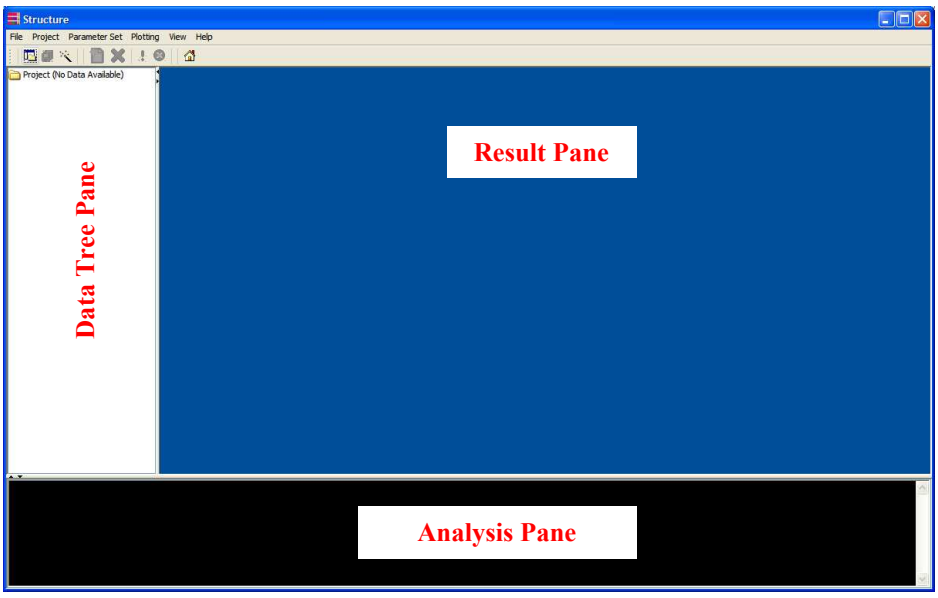
	A	B	C	D	E	F	G	H	I	J	K
1		SSR1	SSR2	SSR3	SSR4	SSR5	SSR6	SSR7	SSR8	SSR9	SSR10
2	GENO1	110	330	190	140	220	140	240	160	200	180
3	GENO1	110	330	190	140	220	140	240	160	200	180
4	GENO2	110	330	190	140	230	140	240	160	190	180
5	GENO2	110	330	190	140	230	140	240	160	190	180
6	GENO3	110	320	190	140	220	140	240	160	200	180
7	GENO3	110	320	190	140	220	140	240	160	200	180
8	GENO4	110	320	-999	140	220	140	240	160	200	180
9	GENO4	110	320	-999	140	220	140	240	160	200	180
10	GENO5	110	330	180	140	220	140	240	160	200	180
11	GENO5	110	330	180	140	220	140	240	160	200	180

SSR is the code for markers; GENO is for genotype; -999: missing data value

- Save the data file in Text (tab delimited) type with a suitable filename <genodata.txt>.

B. Download and install STRUCTURE. Latest version of STRUCTURE is available for download at http://pritch.bsd.uchicago.edu/structure.html.

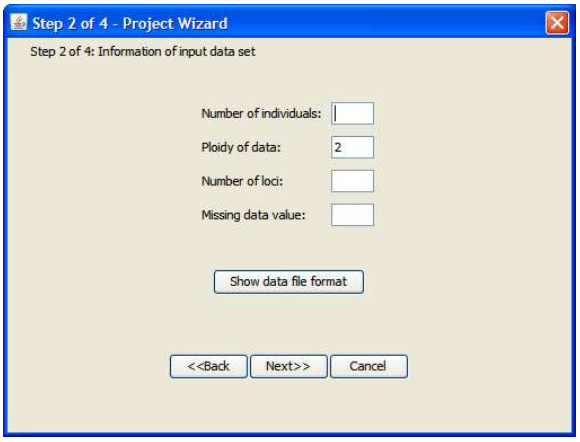
- Run the Structure 3.2.2 software, by double clicking the icon at desktop.



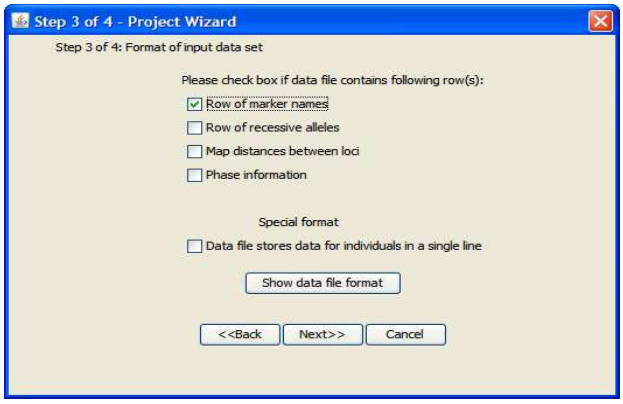
(1) Building a project

- Click on **File > New Project**

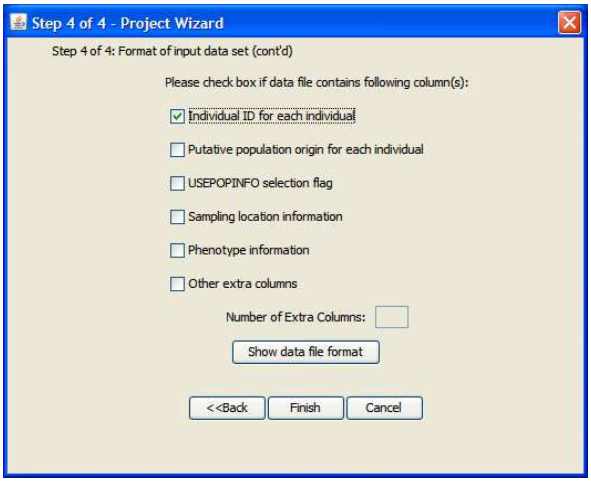
- Fill in these boxes: Name of project, Select directory and Choose data file
- Select the file saved in step A and Click **Next**



- Fill in these boxes: number of individuals, ploidy of data ('2' for diploid), number of loci, and missing data value ('-999'). Click [**Next]**

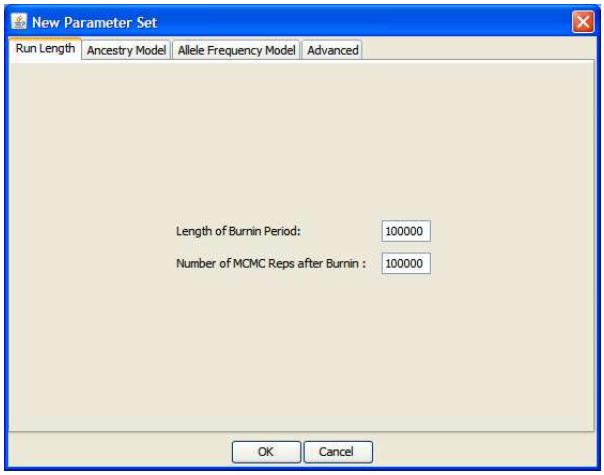


- Since our data contains marker and genotype labels, check 'Row and marker names'. Click [**Next]**

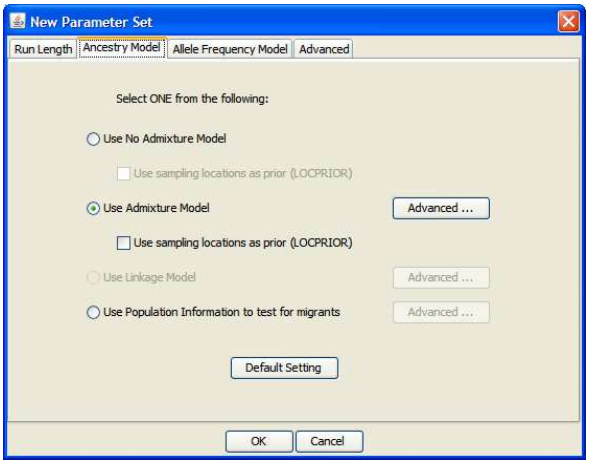


- Since the data file contains genotypes labels, check **Individual ID for each Individual.** Click [**Finish]**

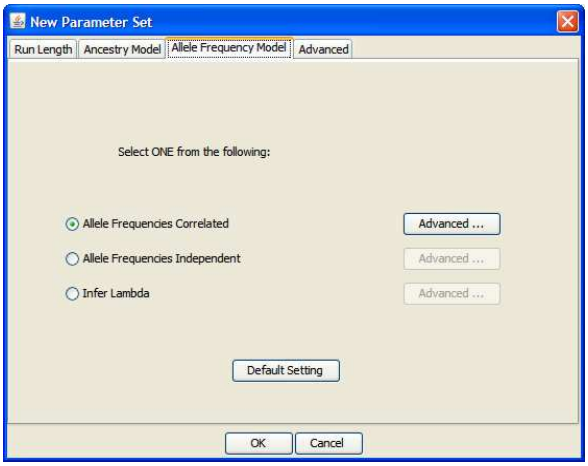
(2) When project is done, a parameter set needs to be configured. For this, in the STRUCTURE Main window, click on **Parameter Set > New**



- Fill in these boxes: Length of Burnin Period: (100000), and Number of Markov chain Monte Carlo (MCMC) Reps (simulations) after Burnin: (100000).
  This number should be high, preferably more than 100000 to get reliable convergence.
- Click **[OK]** button
- In the **Ancestry Model** tab, select Use Admixture Model (This is the default). Click **[OK]** button.



- In the Allele Frequency Model tab, select **Alleles Frequencies Correlated**



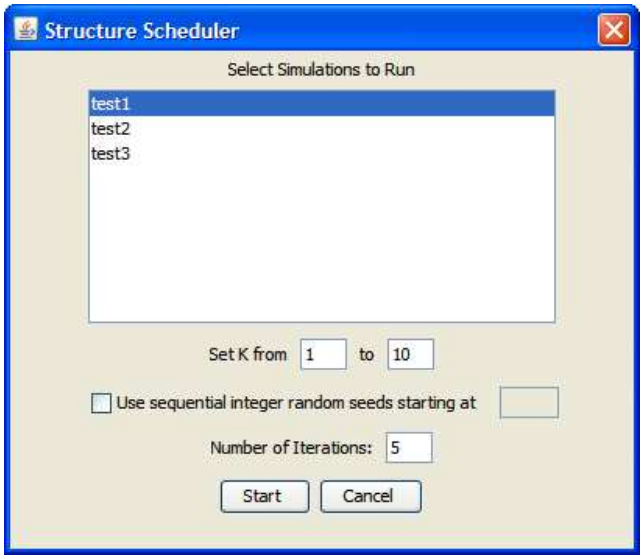
- Click **[OK]** button.

- Name the newly created parameter set in the input dialogue (e.g. test1)
- Click **[OK]**.

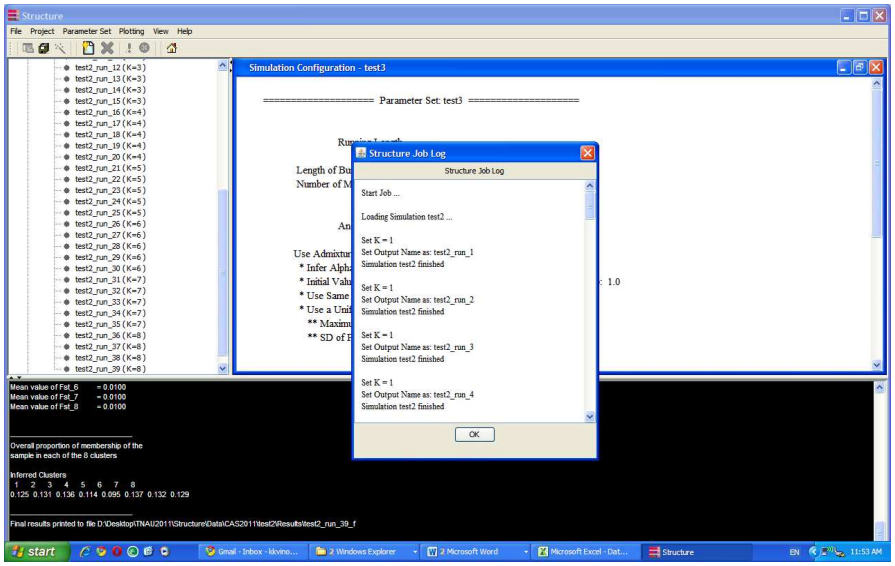
(3) Running Simulations

[If optimum population structure (K) is already known by some other means, then skip this step and go to step (5)]

- In the STRUCTURE main window, click on Project > Start a job



- In the Scheduler dialogue, select the Parameter set you want to run (e.g. test1);
- Set K between a range say 1 to 10;
- Set the number of replications (Iterations) to run (e.g. 5)
- Click **[Start]**

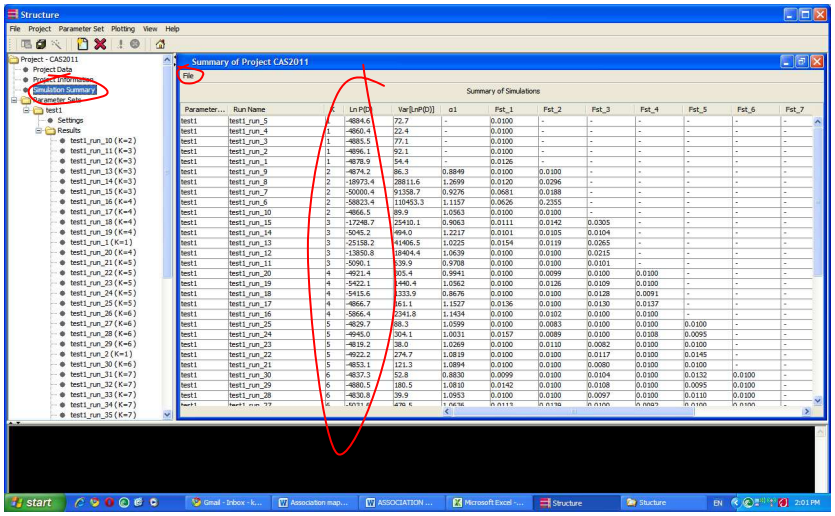


- The program will run for a very long time (>72 hours) depending on the speed of the computer, size of the data, and number of iterations and the replications defined in the parameter set.

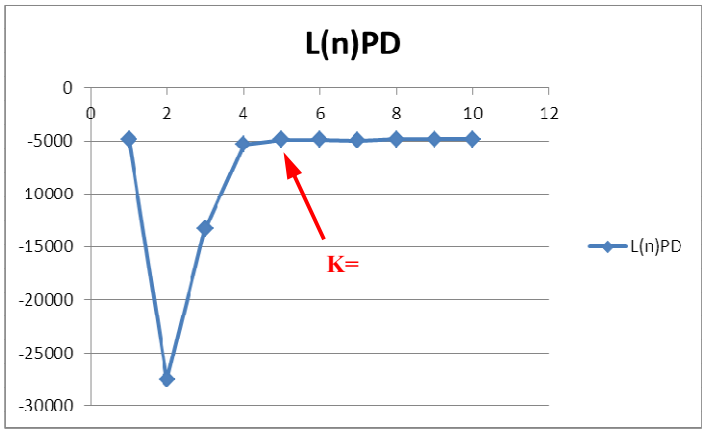
- STRUCTURE displays Job is Completed! dialogue after successful analysis.

(4) Determining the optimum population structure

- To determine optimum value for K, Click on the Simulation Summary in the Tree Pane of STRUCTURE window

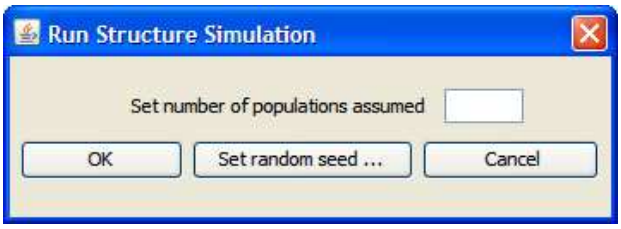


- In the Result pane, click File on the left hand top corner, to save the simulation summary in a text file.
- Copy the values of K and Ln P(D) into a convenient data editor (Excel is the best option) and calculate average of Ln P(D) against each K, across replications. The K at which Ln P(D) plateaus is to be taken as optimum K.



(5) Once optimum population structure (K) is known, estimate Inferred ancestry (Q matrix) of individuals,

- In the STRUCTURE main window, Select **Parameter Set >Run**



- Enter the value of K in the box and click **[OK].**
- When the job is complete, go to the result pane and select the latest run from the results folder
- From the results on the right pane, select inferred ancestry of individuals
- Copy and paste it in notepad.

Alternately,

- Go to the directory in where you save the project, open the folder with project name, and open the result folder. There will be several files with a "f" suffix.
- Open the file with later run number in Notepad. In this output file, copy the values of "Inferred ancestry of individuals"
- Inferred ancestry of individuals (Q matrix) is used as covariate in TASSEL.

- A typical Q matrix will look like as follows:

1	GENO1	0 :	0.212	0.020	0.131	0.043	0.593
2	GENO2	0 :	0.173	0.201	0.538	0.062	0.027
3	GENO3	0 :	0.200	0.270	0.131	0.189	0.211
4	GENO4	0 :	0.092	0.506	0.155	0.142	0.105
5	GENO5	0 :	0.124	0.329	0.046	0.427	0.074
6	GENO6	0 :	0.339	0.053	0.096	0.450	0.062
7	GENO7	0 :	0.343	0.039	0.246	0.120	0.251
8	GENO8	0 :	0.376	0.208	0.201	0.059	0.155
9	GENO9	0 :	0.172	0.044	0.590	0.137	0.058
10	GENO10	0 :	0.172	0.163	0.131	0.445	0.089
11	GENO11	0 :	0.093	0.470	0.101	0.165	0.171
12	GENO12	0 :	0.313	0.156	0.237	0.108	0.187
13	GENO13	0 :	0.184	0.371	0.299	0.030	0.117
14	GENO14	0 :	0.078	0.159	0.036	0.675	0.052
15	GENO15	0 :	0.705	0.076	0.065	0.078	0.077

- Save the Q matrix. This need to be formatted to be read in TASSEL

II. Association Analysis using TASSEL

Association mapping can produce spurious association between marker and phenotype; therefore, the population structure is an important component in estimating marker – trait associations. This is done by incorporating the Q matrix of inferred ancestry coefficients of the individuals across the sub-populations as covariate in the association mapping analysis.

To refine the results, the kinship coefficients are also used in association analysis. Kinship matrix (K matrix) can be estimated using software such as SPAGeDi (Hardy and Vekemans, 2002) or can be estimated within TASSEL itself.

Unlike that of STRUCTURE which is a complete program by itself, TASSEL stand-alone version runs only under Java runtime environment (JRE) version 1.5 and above. JRE is freely downloadable software from Sun Microsystems, http://java.sun.com/. Alternatively, online versions of TASSEL are also available.

Note: Latest version of TASSEL 3.0 does not support microsatellite data anymore. So SSR data analysis can be done only using TASSEL version 2.1.

Both TASSEL 2.1 and 3.0 are available for free download at the following website:
http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119.

- Once software platforms are ready, double clicking on the file sTASSEL.jar will run TASSEL 2.1.

A. Preparation of data

TASSEL requires three types of data primarily for the analysis. (i) Marker segregation data (ii) Phenotype data and (iii) Ancestry coefficient data (Q matrix)

- Prepare these data in Excel, and save as Text Tab delimited (*.txt) files.

(a) Genotype data

Genotype data using microsatellites uses the following format:

	A	B	C	D	E	F	G	H	I	J
1	40	96:2								
2		SSR1	SSR2	SSR3	SSR4	SSR5	SSR6	SSR7	SSR8	SSR9
3	GENO1	110:110	330:330	190:190	140:140	220:220	140:140	240:240	160:160	200:200
4	GENO2	110:110	330:330	190:190	140:140	230:230	140:140	240:240	160:160	190:190
5	GENO3	110:110	320:320	190:190	140:140	220:220	140:140	240:240	160:160	200:200
6	GENO4	110:110	320:320	190:190	140:140	220:220	140:140	240:240	160:160	200:200
7	GENO5	110:110	330:330	180:180	140:140	220:220	140:140	240:240	160:160	200:200
8	GENO6	110:110	?:?	180:180	140:140	230:230	140:140	240:240	160:160	190:190
9	GENO7	110:110	330:330	190:190	140:140	220:220	140:140	240:240	160:160	200:200
10	GENO8	110:110	320:320	180:180	140:140	220:220	140:140	240:240	160:160	200:200
11	GENO9	110:110	330:330	190:190	140:140	220:220	140:140	250:250	160:160	200:200
12	GENO10	120:120	320:320	180:180	140:140	220:220	140:140	240:240	160:160	200:200

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of markers, and (:2) indicate diploid nature of the individuals. ? is commonly used for missing data. *Don't put individual with missing data in the first row. Instead, move it into another row.*

Genotype data using SNP uses the following format:

	A	B	C	D	E	F	G	H	I	J
1	40	96:2								
2		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
3	GENO1	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
4	GENO2	C:C	G:G	T:T	G:G	A:A	G:G	G:G	A:A	T:T
5	GENO3	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
6	GENO4	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
7	GENO5	C:C	G:G	A:A	G:G	C:C	?:?	G:G	A:A	A:T
8	GENO6	C:C	G:G	A:A	G:G	A:A	G:G	G:G	A:A	T:T
9	GENO7	C:C	G:G	T:T	G:G	C:C	G:G	G:G	A:A	T:T
10	GENO8	C:C	G:G	A:A	G:G	C:C	G:G	G:G	A:A	T:T

11	GENO9	C:C	G:G	T:T	G:G	C:C	G:G	T:T	A:A	T:T
12	GENO10	T:T	G:G	A:A	G:G	C:C	G:G	G:G	A:A	T:T

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of markers, and (:2) indicate diploid nature of the individuals. ? is commonly used for missing data. *Don't put individual with missing data in the first row. Instead, move it into another row.*

- Save the data matrix in Text (Tab delimited) type with a suitable filename, <Markername.txt>.

(b) Phenotype data

Phenotype data uses following format:

	A	B	C	D	E	F	G
1	40	6	1				
2		PHE1	PHE2	PHE3	PHE4	PHE5	PHE6
3	GENO1	12.72	21.64	121.23	88.30	2.40	12.72
4	GENO2	11.32	25.16	129.20	95.30	2.46	11.32
5	GENO3	12.38	25.32	139.10	92.35	2.72	12.38
6	GENO4	13.00	25.19	123.60	104.80	2.26	13.00
7	GENO5	12.67	24.19	129.70	97.50	2.95	12.67
8	GENO6	10.80	24.24	118.10	86.10	2.57	10.80
9	GENO7	9.62	27.92	129.60	94.85	1.94	9.62
10	GENO8	9.35	25.30	114.20	96.70	2.28	9.35
11	GENO9	9.68	25.41	83.70	99.70	1.65	9.68
12	GENO10	9.16	26.44	94.50	91.00	2.10	9.16

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of traits, and 1 indicate number of header rows. -999 is commonly used for missing data.

- Save the phenotype data matrix in Text (Tab delimited) type with a suitable filename, <traitname.txt>.

(c) Population structure data

Population structure data (Q matrix) uses following format:

	A	B	C	D	E	F
1	40	5	1			
2		Q1	Q2	Q3	Q4	Q5
3	GENO1	0.000	0.003	0.037	0.003	0.956
4	GENO2	0.000	0.003	0.006	0.016	0.975
5	GENO3	0.000	0.001	0.001	0.001	0.996
6	GENO4	0.000	0.004	0.005	0.002	0.989
7	GENO5	0.000	0.001	0.001	0.001	0.996

8	GENO6	0.000	0.001	0.002	0.001	0.996
9	GENO7	0.001	0.003	0.629	0.004	0.362
10	GENO8	0.000	0.002	0.001	0.001	0.995
11	GENO9	0.000	0.021	0.004	0.125	0.850
12	GENO10	0.001	0.002	0.002	0.002	0.993

Note: The number in the first row tell TASSEL, the number of individuals, followed by number of sub-populations (K=5), and 1 indicate number of header rows.

- Save the Q matrix in Text (Tab delimited) type with a suitable filename, <Q_matrix name.txt>.

(d) Kinship data (K matrix)

Kinship data is an optional requirement for Association mapping. Structured association analysis is done using a general linear model (GLM) algorithm, which does not require K matrix. K matrix is however, essential for mixed linear model (MLM) analysis.

If the kinship output from SPAGeDi is used, it should be formatted to read in TASSEL as given below. For this, (i) add a value of "2" for relative kinship between same individuals and (ii) change the all negative values of relative kinship into "0".

	A	B	C	D	E	F
1	40					
2	GENO1	2.000	0.595	1.688	1.688	0.506
3	GENO2	0.595	2.000	0.572	0.550	1.286
4	GENO3	1.688	0.572	2.000	1.465	0.483
5	GENO4	1.688	0.550	1.465	2.000	0.416
6	GENO5	0.506	1.286	0.483	0.416	2.000

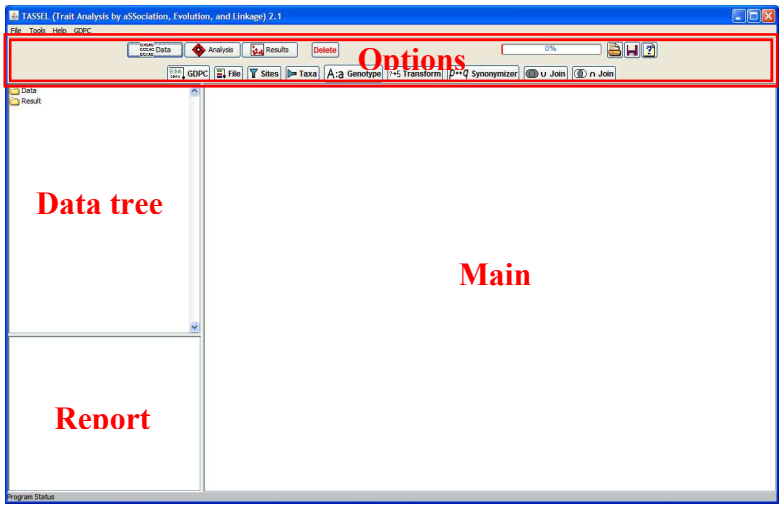
Note: The number in the first row tell TASSEL, the number of individuals. No missing data are permitted in K matrix.

- Save the K matrix in Text (Tab delimited) type with a suitable filename <kinship.txt>

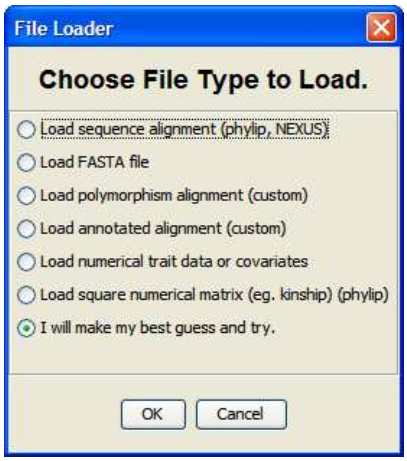
B. Running Structured Association Mapping

(i) Loading data

- Double click on the file sTASSEL.jar in the TASSEL 2-1 directory. Following window opens.



- Click on the [**Data]** button from Options panel.
- From the buttons below click on [**File]** button

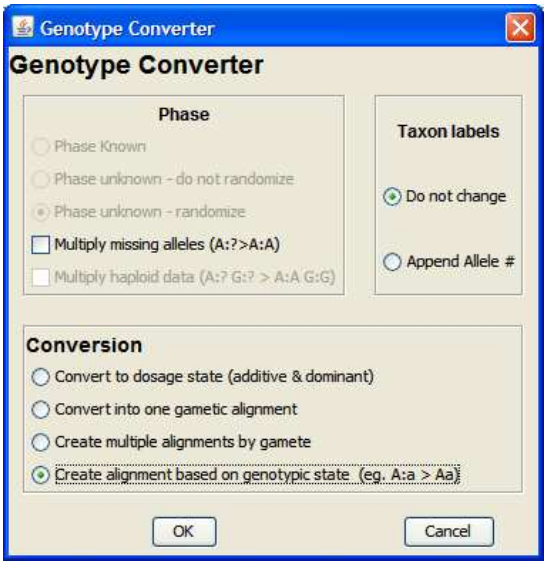


- Select the data type to load, if not sure about data, it is better to choose "I will make my best guess and try", this will allow TASSEL to select the data type by itself.
- Click **[OK]** and select the file containing marker name <markername.txt>.
- Repeat the step, and load phenotype data <traitname.txt> and Q-matrix <Q_matrixname.txt> one after the other.
- Once the data are loaded, they appear in the Data tree panel.

(ii) Geneotype data processing

When diploid microsatellite data is used, convert the raw format to genetype state, to do this,

- Click on the button [**A:a Genotype]** to start Genetype Converter



- Select **Create alignment based on genotypic state (eg. A:a > Aa)** and click **[OK]**
- This will add another dataset named "GenoStates" in the Data tree panel

(iii) Joining marker, phenotype and population structure data

- In the Data tree Panel, select data "GenoStates", "<Traitname>" and "<Q_Matrixname>" by clicking on them, while <Ctrl> key is pressed
- Click on the button **[U Join]** in the Options Panel
- A new Data set named "GenoStates+<Traitname>+<Q_matrixname>" appear on the data tree panel

(iv) Loading relative kinship data (for MLM analysis only)

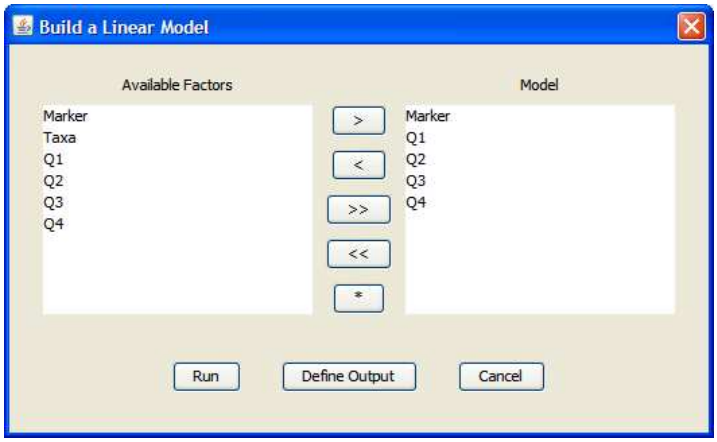
- If kinship information is available, load it by clicking **[File]** button on the Options Panel.
- Select either "Load square numerical matrix (eg. kinship) (phylip)" or "I will make my best guess and try" and click **[OK]**
- Select Kinship file and Open
- The kinship data appear under Matrix in the Data tree panel
- Alternately Kinship can be calculated within TASSEL, by selecting the "GenoStates" and clicking on **[Analysis]** and then **[Kinship]**.

  Note: This is a simple kinship matrix generated from the distance matrix. In order to use more robust Kinship estimates it is recommended to use SPAGeDi or SAS.

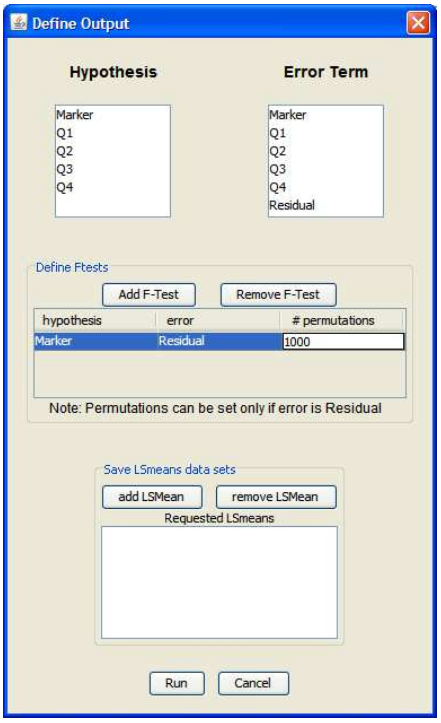
(v) Structured association analysis using least squares GLM

- Select "GenoStates+<Traitname>+<Q_matrixname>" from the Data tree panel, by clicking on it while holding the <Ctrl> key pressed
- Click on the **[Analysis]** button and then click on **[GLM]**

- Select all phenotype as data, and Sub-population data as covariate, Exclude the last sub-population
- Check "Analyse Each Data Column Separately"
- Click **[OK]**



- Click on **[Define Output]**



- In Define Ftests, we can set number of permutations, say 1000. Click **[Run]**
- A new data set appear under Result>Association in the Data tree Panel named "GLM_ GenoStates+<Traitname>+<Q_matrixname>"

(vi) Viewing and saving results

- Click on the button [Results] from Options Panel
- Select the result data from Data tree Panel, by holding the <Ctrl> key pressed and clicking on "GLM_ GenoStates+<Traitname>+<Q_matrixname>"
- Click [Table] button from the Options Panel

Fixed Effect Linear Model Solution for Phenotypes and Genotypes

Trait	Locus	Site	Chr	Chr_pos	df_Ma...	F_Marker	p_Marker	#perm...	p-per...	p-adj_...	df_Model	df_Error	MS_Error	Rsq_M...	Rsq_M...
SPY	RM1	0	0	0	1	0.0187	0.8921	3000	0.8957	1	5	33	7.6772	0.1557	4.7808E-4
SPY	RM101	0	0	0	1	0.0612	0.8061	3000	0.8137	1	5	33	7.6674	0.1568	0.0016
SPY	RM107	0	0	0	1	8.1549	0.0074	3000	0.0117	0.0643	5	33	6.1595	0.3226	0.1674
SPY	RM11	0	0	0	1	0.2542	0.6175	3000	0.6148	1	5	33	7.6229	0.1617	0.0065
SPY	RM127	0	0	0	2	0.5874	0.5616	3000	0.5648	1	6	32	7.6411	0.1851	0.0299
SPY	RM13	0	0	0	1	2.0689	0.1597	3000	0.1799	1	5	33	7.2284	0.2051	0.0498
SPY	RM144	0	0	0	1	0.0401	0.8425	3000	0.8414	1	5	33	7.6723	0.1563	0.001
SPY	RM152	0	0	0	1	1.2263	0.2761	3000	0.2832	1	5	33	7.4064	0.1855	0.0303
SPY	RM153	0	0	0	2	0.2129	0.8094	3000	0.8177	1	6	32	7.8176	0.1663	0.0111
SPY	RM154	0	0	0	2	1.3971	0.262	3000	0.2602	1	6	32	7.2855	0.2231	0.0678
SPY	RM16	0	0	0	2	0.2159	0.807	3000	0.8117	1	6	32	7.8162	0.1665	0.0112
SPY	RM168	0	0	0	1	0.0028	0.9583	3000	0.955	1	5	33	7.6809	0.1553	7.0974E-5
SPY	RM169	0	0	0	3	0.437	0.7281	3000	0.7358	1	7	31	7.8454	0.1895	0.0343
SPY	RM17	0	0	0	1	0.1233	0.7278	3000	0.7354	1	5	33	7.653	0.1584	0.0031
SPY	RM170	0	0	0	3	0.9776	0.4159	3000	0.4385	1	7	31	7.4704	0.2282	0.073
SPY	RM171	0	0	0	1	3.0403E-4	0.9862	3000	0.9907	1	5	33	7.6815	0.1552	7.7828E-6
SPY	RM18	0	0	0	2	0.037	0.9637	3000	0.9687	1	6	32	7.9034	0.1572	0.0019
SPY	RM182	0	0	0	1	2.2052	0.147	3000	0.1653	1	5	33	7.2004	0.2081	0.0529
SPY	RM184	0	0	0	1	2.3627	0.1338	3000	0.1306	1	5	33	7.1684	0.2117	0.0564

Print Export (CSV) Export (Tab)

- By clicking on the **[Print]** results can now be printed, or exported to Tab delimited text file or Comma separated values (CSV) text file by clicking on buttons **[Export (CSV)]** and **[Export (Tab)]** respectively.

(vii) Understanding the result file

Trait	Locus	Site	Chr	Chr_pos	df_Marker	F_Marker	p_Marker	#perm_Marker	p-perm_Marker	p-adj_Marker	df_Model	df_Error	MS_Error	Rsq_Model	Rsq_Marker
SPY	RM1	0	0	0	1	0.0187	0.8921	3000	0.8957	1	5	33	7.6772	0.1557	4.78E-04
SPY	RM101	0	0	0	1	0.0612	0.8061	3000	0.8137	1	5	33	7.6674	0.1568	0.0016
SPY	RM107	0	0	0	1	8.1549	0.0074	3000	0.0117	0.0643	5	33	6.1595	0.3226	0.1674
SPY	RM11	0	0	0	1	0.2542	0.6175	3000	0.6148	1	5	33	7.6229	0.1617	0.0065
SPY	RM127	0	0	0	2	0.5874	0.5616	3000	0.5648	1	6	32	7.6411	0.1851	0.0299
SPY	RM13	0	0	0	1	2.0689	0.1597	3000	0.1799	1	5	33	7.2284	0.2051	0.0498
SPY	RM144	0	0	0	1	0.0401	0.8425	3000	0.8414	1	5	33	7.6723	0.1563	0.001
SPY	RM152	0	0	0	1	1.2263	0.2761	3000	0.2832	1	5	33	7.4064	0.1855	0.0303
SPY	RM153	0	0	0	2	0.2129	0.8094	3000	0.8177	1	6	32	7.8176	0.1663	0.0111
SPY	RM154	0	0	0	2	1.3971	0.262	3000	0.2602	1	6	32	7.2855	0.2231	0.0678
SPY	RM16	0	0	0	2	0.2159	0.807	3000	0.8117	1	6	32	7.8162	0.1665	0.0112

The result file, in addition to displaying the F-statistics and p-values for the requested F-tests, also contains information about degrees of freedom, the error mean square for the model, R-square of the model, and Rsquare for the marker. The model R-square is the portion of total variation explained by the full model. The marker R-square is the portion of total variation explained by the marker but not by the other terms in the model. When permutations are requested, #perm_Marker is the number of permutations run, pperm_ Marker is a test of individual markers, and p-adj_Marker is the marker p-value adjusted for multiple tests. The p-adj_Marker value is a permutation test derived using a step-down MinP procedure (Ge et al. 2003) and controls the family-wise error rate (FWER). For example, if only markers with p-adj values of .05 or less are accepted as significant, then the probability of rejecting a single true null hypothesis across the entire set of hypotheses is held to .05 or less. This test takes dependence between hypotheses into account and does not assume that hypotheses are independent as do other multiple test correction procedures.

Note:

Both STRUCTURE and TASSEL comes with well written tutorials. This document is no substitution for those. For any clarification and in depth information please read these tutorials carefully. Besides, there are online discussion forums available for these software packages, in

which users post their doubts and suggestions. These discussions are watched by the developers of these software and they incorporate modifications/ fix bugs as and when required.

To join these forums visit following sites,

STRUCTURE: https://groups.google.com/d/forum/structure-software

TASSEL : http://groups.google.com/d/forum/tassel

Major References :

Buckler, E., Casstevens, .T, Bradbury, P., Zhang, Z. 2009. Trait Analysis by aSSociation, Evolution and Linkage (TASSEL): User Manual. Cornell University

http://www.maizegenetics.net/tassel/docs/TASSEL_help.pdf

Pritchard, J.K., Wena, X., Falush, D. 2010. Documentation for structure software: Version 2.3. Department of Human Genetics, University of Chicago.

http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.3/structure_doc.pdf

Other references:

Bradbury, P.J., Zhang, Z. , Kroon, D.E. , Casstevens, T.M., Ramdoss, Y., Buckler, E.S. 2007 TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633-2635

Churchill, G., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., et al. 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet* 36, 1133-1137

Hardy, O.J., Vekemans, X. 2002 SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2,618-620

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. 2000 Association mapping in structured populations. *Am J Human Genet* 67, 170-181.

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000 Inference of population structure using multilocus genotype data. Genetics, 155, 945–959

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., et al. 2001 Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genet* 28, 286-289.

Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., et al. 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet* 38,203-208

2D Gel Electrophoresis Protocol for Proteomics Research

N.Senthil and M. Raveendran, Associate professors,
Department of Plant Molecular Biology and Biotechnology, CPMB, TNAU
senthil_natesan@yahoo.com & raveendrantnau@gmail.com

Introduction

Two-dimensional gel electrophoresis involves the separation of proteins based on itys charge and molecular mass to provide a much greater separation of complex protein mixtures than either of the individual procedures. The most common two-dimensional technique uses isoelectrofocusing (IEF) in a tube gel followed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) in a perpendicular direction. This combination of isoelectric point (pI) and size separation is the most powerful tool for protein separations currently available. After staining, proteins appear on the final two-dimensional gel as round or elliptical spots instead of the rectangular bands observed on one-dimensional gels. Although the total separating power of large-format two-dimensional gels is estimated to be as high as 5000 spots per gel, in practice a single two-dimensional separation of a complex mixture such as a whole-cell or tissue extract may produce 1000 to 2000 well-resolved spots when a sensitive detection method is used.

The most common IEF procedures are based on the use of soluble ampholytes, which are relatively small organic molecules with various isoelectric points and buffering capacities. The pH gradient for IEF gels is produced when the soluble ampholytes migrate in the gel matrix until they reach their isoelectric point. Because stable pH gradients outside the pH 3.0 to 8.0 range are difficult to create, alternative protocols using non-equilibrium conditions are required to resolve proteins with pI values below 3.0 to 4.0 (for acidic proteins) or above 8.0 (for basic proteins). One of the more important limitations of soluble ampholytes is the difficulty in obtaining highly reproducible pH profiles, especially when very narrow pH ranges are needed.

An increasingly attractive alternative to soluble ampholytes is the use of immobilized pH gradient (IPG) strip gels. In this system, the buffering side chains are covalently incorporated into the acrylamide matrix, and any pH range and curve shape can be generated by pouring a gradient gel using two solutions that differ in ampholyte composition rather than acrylamide concentration. As with tube gels, the initial IEF separation is followed by a second separation using SDS-PAGE in a perpendicular direction. The use of IPG gels has recently increased, for at least three major

reasons: many of the technical problems associated with their use have been solved or substantially minimized, reproducible pre-made IPG gels are now commercially available, and lately strong interest has arisen in using two-dimensional gels for proteome analysis studies (analyzing and comparing the complete protein profiles of cell lines, tissue samples, or single-celled organisms).

NOTE: High-purity water (e.g., Milli-Q water or equivalent) and ultra-pure chemicals are essential for all solutions.

Protein Extraction

Leaf tissues were frozen and ground into fine powder with liquid nitrogen. The powder was suspended in a pre-cooled (-20ºC) solution of 10% trichloroacetic acid (TCA) in acetone with 0.07% DTT. The suspension was incubated at -20ºC for one hour and centrifuged (15 minutes, 35000 g at 4ºC). The pellet was dissolved in ice-cold acetone containing 0.07% DTT, incubated at -20ºC for one hour and then centrifuged (15 minutes, 35000 g, 4ºC). (This washing may be continued for two to three times). The resultant supernatant was discarded and the pellet was freeze-dried. Proteins were later solubilized in lysis buffer (9M Urea, 4% (w/v) CHAPS, 0.8% (w/v) Biolyte-Ampholyte pH 3-10, 1% (w/v) DTT). Ten-milligram samples were suspended in 250 µl of lysis buffer and incubated at 37ºC for one hour with continuous stirring and then centrifuged at 10000 g at room temperature. The supernatant serves as the protein extract for analysis. The protein concentrations were measured by the Bradford method (another protocol).

First and Second Dimension Gel Electrophoresis

For both analytical and preparative gels, the 18 cm IPG strips (pH 4-7) were rehydrated overnight with 350 µl of rehydration buffer (8M Urea, 2% CHAPS, DTT (7 mg per 2.5 ml of rehydration buffer) and 0.5% (v/v) IPG buffer pH 4-7) containing the required quantity of proteins in a reswelling tray (APBiotech) at room temperature. For analytical and preparative gels, 100µg and 1.5mg of protein were loaded, respectively. Isoelectric focusing (IEF) was conducted at 20ºC with a Pharmacia Multiphore II kit. The running conditions were as follows: 500V for 1 hour followed by 1000V for 1 hour and finally 3000V for 16 hours. The focused strips were equilibrated twice for 15 minutes in 10ml equilibration solution. The first equilibration was performed in a solution containing 6M urea, 30% (w/v) glycerol, 2% (w/v)

SDS, 1% (w/v) DTT and 50mM Tris-HCl buffer. The second equilibration was performed in a solution modified by the replacement of DTT by 4% (w/v) iodoacetamide. Separation in the second dimension was performed by SDS-PAGE in a vertical slab of acrylamide (12% total monomer, with 2.6% crosslinker) using PROTEAN II MultiCell (BioRad) kit.

Gel Silver Staining

After the termination of the second dimension run, the gels were immersed in fixative solution (methanol/distilled water/acetic acid, 40/50/10) for one hour. The gels were sensitized by exposure to thiosulfate reagent (0.02% Sodium thiosulfate), followed by impregnation with silver nitrate reagent (0.2% silver nitrate and 0.02% of 37 % formaldehyde) for 30 minutes and developed in developing solution (3% sodium carbonate, 0.05% formaldehyde (37%), 0.0005% sodium thiosulfate). The staining reaction was stopped by using 0.5% glycine solution (for 5 minutes) and gels were rinsed with water several times prior to densitometry.

Image and Data Analysis

Silver stained gels were scanned using a GS-800 Densitometer (BioRad) at a resolution of 600 dots and 12-bit per inch. Image treatment, spot detection and protein quantification were done using Melanie 3 software (GeneBio, Geneva, Switzerland). Spot detection was carried out on 12-bit images using optimized parameters as follows: number of smooths, 1; Laplacian threshold, 5; partial threshold, 5; saturation, 90; peakness increase, 100; minimum perimeter, 10. Gel matching was performed by Melanie 3 software and spot pairs were confirmed visually. The scatter plots between gels of each data point were displayed to estimate gel similarity or experimental errors and calibration was performed using fitting report whenever (Melanie 3 user manual). The molecular masses of protein on gels were determined by co-electrophoresis of standard protein markers (APBiotech) and pI of the proteins were determined by migration of the protein spots on 18cm IPG (pH 4-7, linear) strips.

Bioinformatics Tools for Genomics Research

Participants List

S.No	Name	Email Id
1.	Dr.N.M.Arivudainambi, Assistant Professor (Ag.Ento.) Maize Research Station, Vagarai, Palani	maize_ento2rediffmail.com
2.	Dr.D.Vidhya, Assistant Professor (Horti.) Rice Research Station, Ambasamudram,	onlyvidhyaadi@gmail.com dv79@tnau.ac.in
3.	Dr.K.Hemaprabha, Assistant Professor (Biotech.) Horticulture College and Research Institute, Periyakulam	hema.kswamy@gmail.com
4.	Dr.A.Ramalakshmi, Assistant Professor (Ag.Micro.) Dept. of Fruit Crops, Horticulture College and Research Institute, Periyakulam	ramalakshmia@gmail.com
5	Dr.S.Chitra, Assistant Professor (PBG) Department of Spices and Plantation Crops, Horticulture College and Research Institute, Periyakulam	chitrapbg@rediffmail.com
6	Dr. S. Arulselvi, Assistant Professor (PBG) Agricultural Research Station, Vaigai Dam	arulselvisoosai@yahoo.co.in
7	Dr. M.Dhandapani, Assistant Professor(PBG) Regional Research Station, TNAU, Paiyur	dhanda1977@gmail.com
8	Dr.P.Senthilkumar, Assistant Professor (Nemato.) Horticulture Research Station, Yercaud	agrips@rediffmail.com
9.	Dr. S.R.Madhan Shankar, Faculty Dept. of Biotechnology, PSG College of Arts & Science, Coimbatore	srmadhanshankar@gmail.com shankarmadhan@yahoo.co.uk
10.	Dr. A.Gopikrishnan ,Research Associate CPMB,TNAU, Coimbatore	agopikrishnan@yahoo.com
11.	R.Ramjegathesh, SRF CPMB, TNAU, Coimbatore	ramjegathesh@gmail.com
12.	G.Senthilraja, Student Dept. of Plant Pathology, TNAU, Coimbatore.	senthiltnau@yahoo.co.in
13.	G. Rajesha , Student Dept. of Plant Pathology, TNAU, Coimbatore.	rajeshag337@gmail.com
14.	Mr. D.Nagaraja, Research Scholar, Dept. of Biotechnology & Bioinformatics, Kuvempu University, Shankaraghatta, Karnataka	nagarajadk@gmail.com
15.	Ms. C.Priyadharshini, SRF Dept. of Millets, TNAU, Coimbatore	dharshiagri@yahoo.com