J. Hortl. Sci.
Vol. 5 (2): 85-93, 2010

FOCUS

Statistical genomics and bioinformatics

Prem Narain
29278 Glen Oaks Blvd. W.
Farmington Hills, MI 48334 USA
E-mail: narainprem@hotmail.com

ABSTRACT

Some important and interesting topics in the newly emerging disciplines of *Statistical genomics* and *bioinformatics* have been discussed briefly in relation to plants with possible references to fruit crops. This paper is therefore divided into two parts relating to the two disciplines, respectively. In the first part, mapping of quantitative trait loci (QTL), association mapping, mapping of gene expression transcripts (eQTL), marker-assisted selection, and a systems approach to quantitative genetics have been dealt with. In the second part, generation of databases, annotation, annotated sequence databases, and sequence similarity search have been described.

Key words: Statistical genomics, bioinformatics, fruit crops, eQTL, annotated sequence databases, sequence similarity search

I. STATISTICAL GENOMICS

INTRODUCTION

Most of the traits of economic importance in plants have an underlying genetic basis involving several genes, and, are subject to modification by environmental factors. Statistical considerations have been predominant in dissecting such complex traits into estimable components (Narain, 1990). Heritability of a trait, as a proportion of the phenotypic variation that is attributed to genetic causes, has been a prime indicator helpful in taking decisions for genetic improvement of economic traits. Prediction of response to artificial selection (based on intensity and accuracy of selection) and the existence of genetic variability have been successful across several crop plants. However, relationship between the phenotype and the genotype has been like a black box where inferential approach has been the only way to look into it. This scenario is now changing with advent of the modern technologies of gene sequencing, microarray experiments and the enormous advances made in attempts to understand gene and protein expression within the cell of an organism. In this context, information on molecular markers has been extremely helpful in identifying regions on chromosomes (QTL) that bring about variation in a trait, thereby providing tools that can lead to far more accurate selection procedures for genetic improvement. Saturated genetic maps of markers, giving their order along a chromosome and relative distances between them, have been developed. Gene transcript data from microarray experiments can be integrated with molecular marker information to map expression traits (eQTL) that can possibly lead to causal networks. The network approach connecting data on genes, transcripts, proteins, metabolites, etc. indicates emergence of a systems quantitative genetics (Narain, 2009, 2010).

Mapping of Quantitative Trait Loci (QTL)

Genomic techniques like *restriction fragment length polymorphism* (RFLP), *random amplified polymorphic DNA* (RAPD), *amplified fragment length polymorphism* (AFLP), *variable number of tandem repeats* (VNTR) - that consist of micro satellites (short sequences) termed as *short tandem repeats* (STR) or *simple sequence repeats* (SSR) and mini satellites (long sequences) - and *single nucleotide polymorphisms* (SNP) have been developed that help in identification of QTLs by correlation between a trait and its specific DNA markers (Narain, 2000). The first problem is, therefore, to construct a linkage map that indicates the position and relative genetic distances between markers along the chromosomes. Map distance is based on the total number of cross-overs between the two markers, whereas, physical distance between them is denoted in terms of nucleotide base pairs (bp). A centi-Morgan (cM), corresponding to a cross-over of 1%, may span 10 kbs to 1,000 kbs and can vary across species.

Since marker genotypes can be followed for their inheritance through generations, these can serve as molecular tags for following the QTL, provided they are linked to the QTL. This requires detecting the marker-QTL linkage and, if established, estimating the QTL map position on the chromosome. However, these problems depend on whether we have data on experimental populations obtained from controlled crosses, as in plants, or on natural populations like humans where controlled crosses cannot be made.

The most popular method, given by Lander and Botstein (1989), is that of s*imple interval mapping* (SIM). It involves formation of intervals by pairing of adjacent markers and treating them as a single unit of analysis for detection and estimation purposes. It is based on joint frequencies of a pair of adjacent markers and a putative QTL flanked by the two markers. Suppose markers *A* and *B* are linked with recombination fraction *r* and QTL *Q* is located between them with r_1 recombination from *A* and r_2 from *B*. Then, $r = r_1+r_2-2r_1r_2 \cong r_1+r_2$, on the assumption of no interference and *r* so small that no double cross-overs can be assumed. In the classical back-cross design with three loci each with two alleles, *A-a*, *B-b*, and *Q-q*, the expected frequencies for the eight marker-QTL genotypes can be used to obtain conditional probabilities of the QTL genotypes, given the marker genotypes. By setting up a linear regression model between the trait (Y) and the indicator variable (X) taking the value 1 if the QTL is *QQ* and –1 if it is *Qq*, one can estimate a regression coefficient that defines the allelic substitution effect of this QTL. In such a model, the QTL genotype for a given individual is unknown. X is then a *random* indicator variable with conditional probabilities of obtaining *QQ* or *Qq* at the QTL. This means the observed value is modelled as a mixture-distribution with mixture ratios as the *conditional* probabilities. We have, therefore, a situation often referred to as a linear regression with *missing* data. The problem of estimation then involves the use of EM algorithm. By assuming that the character is normally distributed within each of the eight marker-QTL classes with equal variance σ^2, one can set up a likelihood function in terms of unknown parameters, and develop a log likelihood ratio (Λ) for testing the hypothesis that the QTL is not located in the interval where the log likelihoods are evaluated using the maximum likelihood estimates of the genotypic values for the two QTL genotypes, the variance σ^2 and the recombination fraction r_1 between marker *A* and the putative QTL using iterative procedures based on EM algorithm. This statistic is distributed as χ^2 with 1 d.f. The associated lod score for the interval mapping is then (½) ($\log_{10}$e) Λ. This statistic is evaluated at regularly-spaced points; say 1 or 2 cm distance, covering the interval as a function of the presumed QTL position. Repeating this procedure for each interval along the chromosome and plotting the lod score curve against the interval gives a *QTL likelihood map* that presents evidence for the QTL at any position in the genome. Presence of a putative QTL is assumed if lod score exceeds a certain threshold *T* and the maximum of the lod score function in the map gives an estimate of the QTL position and gene effects. Mapping of QTL by interval method is widely used in practice. Analysis is done through the software MAPMAKER/QTL.

Although SIM is the method for QTL mapping most widely used with advantage in several practical situations, it ignores the fact that most quantitative traits are influenced by numerous QTLs. This is overcome either by adopting a model of *Multiple QTL Mapping* (MQM) or by combining SIM with the method of multiple linear regression, a procedure known as *composite interval mapping* (CIM). In all these methods, one uses the approach of maximum likelihood that produces only point estimates of the parameters such as the number of QTLs, their location, and effects. The corresponding confidence intervals are required to be determined separately by re-sampling methods. Further, the correct number of QTLs is difficult to determine using traditional methods. Their incorrect specification leads to distortion of the estimates of locations and effects of QTLs. To address these problems, a *Bayesian* approach is often adopted wherein the joint posterior distribution of all the unknown parameters given their *prior* distributions and the observed data is computed. For details of these various aspects, one can refer Narain (2003a, 2005).

The first application of interval mapping in plant breeding was to an inter-specific backcross in tomato. The parents for the back-cross were the domestic tomato *Lycopersicon esculentum* (*E*) with fruit mass 65 g and a wild South American green-fruited tomoto *L. chmielewskii* (*CL*) with fruit mass 5 g. A total of 237 back-cross plants were assayed for continuously varying characters like fruit mass, soluble-solids concentration and pH, and, 63 RFLP and 20 isozyme markers spaced at approximately 20 cM intervals were selected for QTL mapping. A threshold *T*=2.4, giving a probability of under 5% that even a single false-positive will occur anywhere in the genome, was used. This corresponds approximately to significance level for any single test as 0.001. The resulting QTL likelihood maps revealed multiple QTLs for each trait (6 for fruit weight, 4

for concentration of soluble solids and 5 for fruit pH) and estimated their location to within 20-30 cm.

Fruit crops

Fruit crops differ from most of the agronomic/forest crops in that they have large plant size, long intergeneration period due to their extended juvenile phase, asexual propagation, high heterozygosity and polyploidy. These practice outcrossing and have a long life. They are mostly woody perennials and their products are usually perishable. The major temperate fruit crops belong to *Rosaceae* family. The most important genera of this family are *Prunus, Malus, Pyrus, Fragaria,* and *Rosa*. Important members of the genus *prunus* are peach, cherry, plum, apricot, almond and of the genus *Malus* is apple. They have been slow to respond to new technologies in breeding, until recently. Characters like yield, blooming, harvesting time and fruit quality have been studied with the help of molecular markers in several fruit crops. Long period from seed to fruiting in such crops is a major problem in breeding studies involving crosses. Vegetative reproduction, on the other hand, allows every population to be immortalized and one can study a given character for as many years and in as many different environments as one wants. Interspecies crosses are possible and most of them have small genomes. For instance peach, the best characterized among *Prunus* species, has a haploid genome size of 164 Mbp only. Most of the *Prunus* species are diploid, with 8 pairs of chromosomes whereas, apple and pear are allotetraploid with 17 pairs of chromosomes.

Saturated linkage maps with transferable markers, RFLPs, and microsatellites have been developed to provide basic tools for studies on QTLs and marker-assisted selection in fruit tree breeding. As a result of a European project, a saturated linkage map of 246 markers (235 RFLPs and 11 isozymes) constructed from an F_2 progeny derived from almond (cv. Texas) x peach (cv. Earlygold) cross – termed TxE map- indicated 8 linkage groups (G1 to G8) with a total distance of 491 cm. This led to a *Prunus* reference map with 652 markers and a further set of 13 maps constructed with a sub-set of these markers has enabled genome comparisons among seven *Prunus* diploid species (almond, peach, apricot, cherry, *Prunus ferganensis, Prunus davidiana,* and *Prunus cerasifera*). These have helped establish the position of 28 major genes affecting various agronomic characters in different species of *Prunus* crops (Dirlewanger *et al*., 2004).

The first linkage map in apples was constructed by a European Consortium based on F_1 progeny derived from the cross cv. Prima x cv. Fiesta (FxF map). There were a total of 290 markers consisting of RFLPs, SSRs, isozymes, RAPD etc., distributed over 17 linkage groups. A more saturated map was constructed with the F_1 progeny derived from the cross cv. Fiesta x cv. Discovery (FxD map) using 840 markers that included 129 SSRs. These maps have been helpful in QTL studies on apple. A comparison between apple and *Prunus* maps suggests a high degree of synteny between these two genera. QTLs for blooming, ripening and fruit quality have been found in peach and apple. Some of these QTLs were found to be located in regions of the genome where major genes were earlier mapped. For instance, in peach a major gene responsible for low fruit acidity was in the same region as QTLs affecting fruit quality, a quantitative trait. In apple too, a major gene coding for malic acid content is located in the same region as QTLs for fruit quality.

Various populations of peach x *Prunus davidiana* crosses with different levels of introgression of the *Prunus davidiana* genome into the cultivated peach viz. F_1, F_2 or BC2 were used to discover the positions of respective QTLs. About 13 QTLs explained up to 65% of the total phenotypic variation for powdery mildew resistance in plants exposed to the disease in different times and environments.

Candidate gene approaches have been adopted for finding associations between genes involved in relevant metabolic pathways and major genes or QTLs in fruit trees. Several resistance gene analogs (RGAs) were mapped in *Prunus* that are at similar genomic positions as genes or QTLs which determine ‘sharka‘ resistance in apricot or root-knot nematode resistance in peach and plum.

Linkage Disequilibrium or Association Mapping

The mapping of QTLs in plants based on data collected from pedigrees of populations formed by crossing inbred lines is on a coarser scale, so that a QTL detected is likely to refer to several genes in a chromosomal region. The approach of population-based association mapping that involves linkage disequilibrium (LD) between markers and the genes underlying complex traits leads, on the other hand, to more accurate mapping of genes. The key idea is that a disease mutation assumed to have arisen once on the ancestral haplotype of a single chromosome in past history of the population of interest is passed on from generation to generation, together with markers at tightly linked loci, resulting in LD. The use of this approach in horticultural crops, though widely prevalent in human genetics, is limited.

Advantages of the two approaches can be combined by detecting QTL initially using linkage mapping with moderate number of markers, followed by a second-stage of high-resolution association mapping in QTL regions that capitalizes on a high-density marker map.

Benefits of linkage and association mapping have recently been combined in a single population of maize by adopting a *nested association mapping* (NAM) approach. The maize NAM population was derived by crossing a common reference sequence strain to 25 different maize lines. Individuals resulting from each of the 25 crosses were self-fertilized for four further generations to produce 5,000 NAM recombinant inbred lines (RILs). This population was first used for initial detection of QTL using the linkage mapping approach. Subsequently, within each diverse strain, high-resolution association mapping was adopted with a high-density marker map. It is significant to note that within each RIL, all individuals are genetically nearly identical. This means we can estimate true breeding value of each line far more accurately by averaging phenotypic measurements of a given trait taken on several individuals with the same genotype.

In a recent experiment, genetic architecture of flowering time in *Zea mays* (maize) was dissected using NAM. About 1 million plants were assayed in eight environments to map the QTLs. About 29 to 56 QTLs were found to affect flowering time. These were small-effect QTLs shared among the diverse families. The analysis showed, surprisingly, absence of any single large-effect QTL. Moreover, no evidence was found of epistasis or environmental interactions. Flowering time controls adaptation of plants to their local environment in an out-crossing species like *Zea mays*. A simple, additive genetic model predicting accurately flowering time in this species is, thus, in sharp contrast to that observed in several plant species which practice self-fertilization (Buckler *et al*., 2009).

Mapping QTLs for Gene Expression profile (eQTL)

The advent of DNA chip technology in the form of cDNA and oligonucleotide microarrays has provided huge and complex data-sets on gene expression profiles of different cell lines from various organisms. Such gene expression profiles have recently been combined with linkage analysis, based on QTL mapping, through molecular markers in what has been termed 'genetical genomics' (Jansen and Nap, 2001). Gene expression, in terms of transcript levels, for each individual of a segregating population are phenotypes that are correlated with markers, genotyped for that individual, to identify QTLs and their location on the genome to which the expression trait is linked. Such expression quantitative trait loci (eQTL) studies are similar to traditional multi-trait QTL studies, but with thousands of phenotypes. It is also important to note that, underlying the gene expression differences, there are two types of regulatory sequence variation. One is *cis*-regulatory that affects its own expression and the other is *trans*-acting or protein coding that affects expression of other genes. The first attempt where transcript abundance was used to study the linkage with QTLs was on budding yeast (Brem *et al*, 2002) based on a cross between a laboratory strain and a wild strain, the parents being haploid derivatives. Heritability estimation was based on haploid segregants and the linkage with a marker was tested by partitioning the segregants into two groups, according to marker genotypes, and comparing expression levels between groups, with Wilcoxon-Mann-Whitney test. They found 8 trans-acting loci, each affecting expression of a group of 7 to 94 genes of related function. Since then, several eQTL studies have been published in species like mice, maize, humans, rats and *Arabidopsis thaliana*.

Apart from study of the eQTL in yeast, Foss *et al*. (2007) investigated protein QTL in the *same* population of the yeast using mass spectrometry. Comparison between genetic regulation of proteins and that of the transcripts revealed that loci that influenced protein abundance differed from those that influenced transcript levels, much against expectations.

Marker–Assisted Selection (MAS)

Molecular markers such as those provided by RFLP have not only made it possible to detect and estimate effects of QTLs, but can also be used as a criterion of indirect selection for genetic improvement of a given quantitative trait – a procedure of selection which has come to be known as *marker-assisted selection* (MAS). The underlying basis of MAS is the correlation between a trait and the marker genotype, which gets generated due to linkage disequilibria between the QTL and marker loci. The fact that such information can be integrated with those of artificial selection on individual and/or collateral basis, to increase the efficiency of selection, was demonstrated by the work of Lande and Thompson (1990). They showed that relative efficiency of the selection index, combining phenotypic and molecular information optimally, is a function of heritability (h^2) of the trait and the proportion (p) of additive genetic variance of

the trait that is associated with marker loci. This efficiency is always one when h^2=1, the phenotype being a perfect indicator of its breeding value. But, for a character with low heritability, the efficiency can be substantially high, provided *p* is high. This means the value of maker information can be very great if a larger proportion of additive genetic variance is associated with the markers. Efficiency is maximum when *p*=1 and is (1/*h*), that becomes infinitely large for extremely small *h*. In that case, all of the weight in selection index is put on molecular information. If we select *only* on the basis of marker information, the efficiency, relative to individual selection with the same intensity, would be. This shows that when $p>h^2$, selection based on marker information *alone* would be more efficient than individual phenotypic selection.

Increased efficiency of MAS, however, is accompanied by increased cost involved in sample collection, DNA extraction and typing of individuals in the sample, compared to that involved in taking simple measurements of the trait. Cost reduction for MAS can be achieved in several ways. Marker technologies such as those based on *polymerase chain reaction* (PCR) may reduce the cost of MAS. *Selective genotyping* of the extreme progeny, as advocated by Lander and Botstein (1989), is another way. Yet another way could be to bring in auxiliary information from other traits that are correlated with the main trait, and are cheaper to measure. This idea has been used in the past by several workers to increase the efficiency of individual and family selection itself, by including in the index one or more auxiliary traits in conjunction with the main trait. As a matter of fact, molecular information in MAS is itself a sort of auxiliary information, but obtained at a higher cost. Narain (2003b), therefore, showed how the efficiency of MAS behaved if information on one or more auxiliary traits with the corresponding molecular scores was combined with that on the main trait, in an optimal manner.

Fruit crops

In fruit crops, molecular markers are used for screening and selecting the best seedlings several years before the characters are evaluated in the field. It saves space and time so important in woody perennials. Marker-assisted selection in such crops is, however, mostly based on major genes, since several characters like disease resistance, flower/fruit/nut quality are found to be controlled by major genes that follow a simple inheritance pattern. Markers tightly linked to such genes are searched for early selection. They are primarily used for characters that cannot be evaluated till the plant has reached the adult stage, such as fruit characters or self-incompatible genotypes. For instance, gametophytic self-incompatibility in almond, apricot and cherry is one such trait that is encoded by a highly polymorphic locus (S/s) located in the distal part of G6 linkage group. With determination of the sequences of the polymorphic S-RNase gene at this locus, a number of species-specific and allele-specific DNA markers were discovered that were used for early and more accurate selection of self-incompatibility or self-compatibility alleles. Markers close to the two genes of resistance to root-knot nematodes are used for selection of resistant *Prunus* rootstocks. The resistance gene *Ma/ma* from Myrobalan plum and located on G7 linkage group, and another one from peach cv. Nemared (*Mi/mi*) located on G2 linkage group, have been screened with markers in a search for rootstocks that pyramid both resistance genes in a three-way progeny obtained from peach, almond and Myrobalan plum.

Marker-assisted selection for disease resistance is quite widespread in apple as a means of early selection, and, to pyramid resistance genes.

Systems approach

As we know, the central dogma of molecular biology stipulates that sequence information flows from DNA to RNA to protein but not in the reverse direction. But, Kimchi-Sarfaty *et al* (2007) reported data that indicate that a protein's three-dimensional structure is *not* necessarily determined by its amino acid sequence that has been specified by the DNA sequence. An mRNA, if subjected to translational braking, can generate a protein with a structure different from that specified by the DNA sequence. This has been termed 'translation-dependent folding' (TDF) hypothesis (Newman and Bhat, 2007). Differential gene expression resulting in transcripts as sub-phenotypes could, then, lead to different proteins and give results similar to those obtained in the yeast experiment, as reported by Foss *et al* (2007). Genes and proteins are, therefore, required to be considered simultaneously to unravel the complex molecular circuitry operating within a cell. One has to have a global perspective of genotype-phenotype relationship, instead of individual components like DNA or protein in a cellular system.

It seems the interplay of genotype-phenotype relationship for quantitative variation is not only complex but also needs a closer look at how we view this relationship – whether purely at the DNA-RNA level (as in the reductionist approach) or at the level of cell as a whole

(where DNA-RNA are just parts of the cellular system with other contextual forces present in the micro-environments of the cell, also playing their own important roles). Such situations have also been noticed in agricultural experimentation where a dialectical approach has been advocated (Narain, 2006, 2008). In the grain production process, it is also important to study how this process affects soil health and the ecosystem surrounding the plant, as is studying the effect of inputs on production. In the dialectical approach, this relationship between the plant and its environment is studied both ways – input to output as well as output to input, a sort of feedback. A similar possibility seems to exist in the genotype and phenotype relationship within a cell. The protein as a phenotype is determined by a DNA sequence as the genotype, but the reverse phenomenon of protein affecting the DNA could also take place at the expense of violating central dogma. In fact, studies are on to explore biochemical signaling pathways that regulate function of living cells through regulatory networks having positive and negative feedback loops, though it is unclear how genetics can be incorporated into it. These feedback loops are basically *cybernetic* concepts that are inherent in the dialectical approach. This approach takes into account dynamics of the system over time as well, in which, development is a consequence of opposing forces. This is based on the concept of *contradiction* inherent in the meaning of *dialectics*. Things change because of the action of opposing forces on them, and things remain what they are because of temporary balance of the opposing forces. Opposing forces are seen as contradictory in the sense that each taken separately would have an opposite effect, but their joint action may be different from result of either acting alone. These forces are, however, part of self-regulation and development of the object is regarded as a network of positive and negative feedback loops, incorporation of which (in the genetic context) would violate the central dogma. Genes, transcripts, proteins, metabolites, physical components, etc., can be regarded as 'parts' of the cellular system and the 'whole' is regarded as a relation of these parts that acquire properties by virtue of being parts of a particular whole. As soon as the parts acquire properties by being together, they impart to the whole new properties that are, in turn, reflected in changes in the parts, and so on. Parts and whole, therefore, evolve as a consequence of their relationship, and the relationship itself evolves. Genes are fixed, but their expression-the transcript-is not. At any given moment of time, genes are expressed as per requirement of the cell and through information contained in its DNA. At this moment of time, the cellular system is said to have a particular *state* of the system. At the next moment of time, the same genes may be expressed, but differently, depending upon the then requirement of the cell and based on the feedback, if any, from the system's state at the previous time point, assuming that the process is Markovian. This gives the next state of the system, which might or might not be different from the previous state. And, the process goes on continually, modifying the relationship between different parts of the system based on interactions and feedbacks. It seems that a dialectical approach could provide the clue for understanding how 'parts' of a system and the 'whole' system behave in the context of genetics.

II. BIOINFORMATICS

INTRODUCTION

Genomic research is creating quantities of data at unprecedented scales by looking at either *all* genes in a genome, or *all* transcripts in a cell, or else *all* metabolic processes in a tissue in several species, in general, and in agriculture in particular. Very soon new genomic technologies will enable individual laboratories to generate terabyte or even petabyte scales of data. To handle these data, to make sense of them and render them accessible to biologists, is the task of a newly emerging field of *bioinformatics* existing at the interface of biological and computational sciences - computer based analysis of large biological data sets. The data sets usually pertain to macromolecular sequences (DNA, RNA and protein sequences), protein structures, gene expression profiles and biochemical pathways. It has three components. Firstly, it involves development of databases to store and search data. Secondly, it deals with statistical tools and algorithms to analyze and determine relationships between data sets. Lastly, it involves application of the tools for analysis and interpretation of various types of genomic data. For a brief discussion on these aspects, reference may be made to Narain (2005). Here, we discuss primarily those aspects that relate to plant genomes.

Generation of Databases

DNA sequences stored in databases are of three types: genomic DNA, cDNA and recombinant DNA. Genomic DNA, taken directly from the genome, contains genes in their natural state which, in eukaryotes, include introns, regulatory elements and a large amount of surrounding inter-genic DNA. cDNA is reverse-transcribed from mRNA and corresponds to only expressed parts of the genome, there being no introns. It gives direct access to genes that represent only a small percentage of the entire sequence. Recombinant DNA comes from the laboratory,

being artificial DNA molecules – sequence of vectors such as plasmids, modified viruses and other genetic elements used in the laboratory.

High quality sequence data is generated by performing multiple reads on both DNA strands. Sequence data of lower quality can, however, be generated by single reads – single pass sequencing on a much larger scale, quickly and cheaply. Expressed sequence tags (ESTs) are generated by single-pass sequencing of random clones from cDNA libraries and are used to identify genes in genomic DNA as well as to prepare large clone sets for DNA microarrays. Most RNA sequences are deduced from the corresponding DNA sequences, or, from a cDNA sequence. The latter is more informative due to it being extensively processed during synthesis. For example, introns are spliced out of a primary transcript to generate mature mRNA.

Plant sequence data are generated through (i) whole genome sequencing, (ii) sample sequencing of bacterial artificial chromosomes (BACs), (iii) genome survey sequencing (GSS), and (iv) sequencing of expressed sequence tags (ESTs). An integrated database and suite of analytical tools to organize and interpret these data, has been developed and is known as PlantGDB (*vide* the website http://www.plantgdb.org/).

Annotation

Annotation means obtaining useful biological information (structure and function of genes and other genetic elements) from raw sequence data. Since prokaryotes and eukaryotes differ in their structure and genome organization, their annotations involve different problems. Prokaryotes have high gene-density with virtually no introns, but in eukaryotes, gene-density is low and the genome has greater complexity.

We have two groups of annotation - structural annotation and functional annotation. In the former, we are concerned with finding genes and other genetic elements in genomic DNA. In the latter, we assign functions to the discovered sequences.

Annotated Sequence Databases

The following three repositories and resources for primary sequence data are available where each entry is extensively annotated. They can be accessed freely over the World Wide Web (www).

(i) Gene Bank of the National Centre for Biotechnology Information (NCBI)

(ii) Nucleotide Sequence Database of European Molecular Biology Laboratory (EMBL)

(iii) DNA Databank of Japan (DDBJ).

New sequences can be deposited in any of the databases, since, these exchange data on a daily basis. The main sequence databases have a number of subsidiaries for storage of particular types of sequence data. For example, dbEST is a division of Gen Bank which is used to store *expressed sequence tags* (ESTs). Other divisions of Gen Bank include dbGSS, dbSTS - used to store *sequence tagged sites* (STSs) - and several others.

These large database providers, however, do not give non-redundant and curated records, so that detailed analysis cannot be performed at the resource site by the user. A data- base like PlantGDB, which downloads raw plant genomic data from Gen Bank, overcomes such difficulties and provides curated records with detailed and updated information. It organizes EST sequences into contigs that represent tentative unique genes. They are duly annotated and linked to their respective genomic DNA. The data-base gives the basis for identifying genes common to particular species by integrating a number of bioinformatics tools that help in gene prediction and cross-species comparison - the goal of comparative genomics.

Besides PlantGDB database, there are species-specific databases like The *Arabidopsis* Information Resource (TAIR), MaizeGDB, Gramene, a tool for grass genomics, and the Stanford Microarray Database. The PlantGDB genome browsing capabilities for *Arabidopsis* are made possible by *A. thaliana* Genome Database (AtGDB; http://www.plantgdb.org/AtGDB/). This database stores EST and cDNA spliced alignments along with current *Arabidopsis* genome annotation.

As we know *Arabidopsis thaliana*, which is a small mustard species – *eukaryotic* and self-pollinating – is already playing an important role as a model organism in development of plant molecular biology, by way of providing increased knowledge and understanding of the plant's functional and developmental processes. It has a rapid life cycle and can be easily grown in laboratory in large numbers. Its entire genome, that is highly compact and consists of about 130 Mb with little interspersed repetitive DNA, has been sequenced. Many thousands of A*rabidopsis* plants can be grown on a bench to search for particular mutants which can then be isolated and genes cloned for use in other crops. It is related to many food plants like rice, wheat, maize, sorghum, millets, etc., and can, therefore, provide a

focus from which genome content of other higher plants can be extrapolated.

Fruit crops

In regard to horticultural crops, an international consortium led by Albert Abbott at Clemson University (Clemson, SC), developed databases on *Prunus* genome. Using RFLPs on the TxE map and a BAC library of peach cv. Nemared, a physical map was assembled. A growing collection of ESTs from peach and almond, based on cDNA libraries, was released to public databases and more than 3,800 peach putative unigenes were detected. About 2,000 of these unigenes were assigned to specific BAC that contain them. Recently, a Rosaceae database (www.genome.clemson.edu/gdr) has been developed that includes apple, peach, cherry, plum, apricot, pear, etc.

Sequence Similarity Searches

Due to molecular evolution, macromolecule sequences share a common ancestor resulting in similarity in their sequences, structure and biological functions. On the other hand, any pair of sequences will share a certain degree of similarity, due to chance alone. For example, DNA sequences are constructed from an alphabet of only four letters, viz., A, T G and C. Any sequence that consists of a mixture of these letters will show some similarity to any other similarly-constructed sequence. We need to distinguish between such a chance similarity and similarity resulting from real evolutionary and/or functional relationship. This requires use of appropriate statistical methods.

Sequences are first aligned in terms of their letters. When identical letters get aligned, we say that these letters were part of the ancestral sequence and have remained unchanged. When non-identical letters get aligned, we say that a mutation has occurred in one of the sequences. It may also happen that some letters in a particular sequence lack an equivalent in the other sequence, resulting in a gap. This could be due to insertion or deletion of letter/s in one of the sequences, with respect to the ancestral sequence. Dynamic programming algorithms – computational methods - can calculate the best alignment of two sequences. The algorithm takes two input sequences and produces the best alignment between them as the output. Well-known algorithms are Smith-Waterman algorithm (local alignment) and Needleman-Wunsch algorithm (global alignment).

To quantify similarity, a simple alignment score measures the number or proportion of identically matching residues. Gap penalties are subtracted from such scores to ensure that alignment algorithms produce biologically sensible alignments, without too many gaps. Gap penalties may be constant, i.e., independent of the length of the gap or be proportional to the length of the gap, or else may be affine, i.e., containing gap-opening and gap-extension contributions.

We have often a query sequence about which we need to predict the structure and/or the function. We perform sequence similarity searches of databases in which the query sequence is aligned (compared) to each database sequence in turn and then rank the database sequences with the highest scoring (most similar) at the top. This can be achieved by the dynamic programming method with Smith-Waterman algorithm but the procedure is very slow, taking hours, for searching large databases. On the other hand, algorithms like BLAST (Best Local Alignment Search Tool) and FASTA provide very fast (about five to fifty times faster) searches of sequence databases. They are however less accurate than the dynamic programming method which provides the best possible alignment to each database sequence. Each of the BLAST and FASTA operates by first locating short stretches of identically or near-identically matching letters (words) –assumed to lead to high scoring alignment - that are eventually extended into longer alignments.

Acknowledgements

This work was supported by the Indian National Science Academy, New Delhi, under their programme “INSA Honorary Scientist”.

REFERENCES

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science, **296**: 752-755

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Rosas, M.O., Rocheford, T.R., Romaij, M.C., Romero, S., Salvo, S., Villeda, H.S., da Silva, H.S., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S. and McMullen, M.D. 2009. The genetic architecture of maize flowering time. Science, **325**: 714-718

Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Caldere, F., Cosson, P., Howad, W. and Arus, P. 2004. Comparative mapping and marker-assisted selection

in Rosaceae fruit crops. *PNAS,* **101**: 9891-9896

Foss, E.J., Radulovic, D., Shaffer, S.A., Ruderfer, D.M., Bedalov, A., Goodlett, D.R., and Kruglyak, L. 2007. Genetic basis of proteome variation in yeast. *Nat. Genet*., **39**: 1369-1375

Jansen, R.C. and Nap, Jan-Peter. 2001. Genetical genomics: the added value from segregation. Trends in Genetics, **17**: 388-391

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V. and Gottesman, M.M. 2007. A "silent" polymorphism in the MDRI gene changes substrate specificity. Science, **315**: 525-528

Lande, R. and Thompson, R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics,* **124**: 743-756

Lander, E.S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics,* **121**: 185-199

Narain, P. 1990. *Statistical Genetics*. New York: John Wiley and Wiley Eastern Ltd., New Delhi. Reprinted in 1993. Published by the New Age International Pvt. Ltd., New Delhi in 1999. Reprinted in 2008

Narain, P. 2000. Genetic diversity – conservation and assessment. *Curr. Sci.,* **79**:170-175

Narain, P. 2003a. Evolutionary genetics and statistical genomics of quantitative characters. *Proc. Ind. Natl. Sci. Acad.,* **B69**:273-352

Narain, P. 2003b. Accuracy of marker-assisted selection with auxiliary traits. *J. Biosci.,* **28**:569-579

Narain, P. 2005. Mapping of Quantitative Trait Loci. *The Mathematics Student,* **74**:7-18, Printed in 2007

Narain, P. 2006. Statistical Tools in Bioinformatics. *The Mathematics Student,* **75**:17-27, Printed in 2007

Narain, P. 2006. Dialectical agriculture. *Natl. Acad. Sci. Lett.*, **29**:253-260

Narain, P. 2008. Dialectical approach to agriculture. *Proc. Indian Natn. Sci. Acad.*, **74**:61-66

Narain, P. 2009. The Genetic Architecture of Quantitative Variation. *Natl. Acad. Sci. Lett.,* **32**:135-1437

Narain, P. 2010. Quantitative Genetics: past and present. *Mol. Breeding,* **26**:135-143

Newman, S. A. and Bhat, R. 2007. Genes and proteins: Dogmas in decline. *J.Biosci*., **32**:1041-1043