1. Introduction

The complexity of problems in statistical and computational science applicable to genetics and molecular biology has successfully attracted many bright statisticians, mathematicians, and computer engineers, who have time and again came out with useful software solutions, computational algorithms, database theories and data mining approaches. In fact, this has been one of the principal factors behind several success stories we have witnessed in life sciences in the last few decades, exemplified by the phenomenal recent achievement in producing the draft sequence of the entire human genome (Venter et al., 2001; International Human Genome Sequencing Consortium, 2001).

There exists a gradient of biotechnologies varying in degree of sophistication, complexity, stage of development, and most importantly, application. At one end of the gradient, there are *in vitro* technologies, fermentation and biofertilizer technologies, and on the other end of the gradient, there are more sophisticated technologies such as genetic engineering, molecular breeding and genomics. In this article, I shall aim to focus only on selected applications of statistical and computational tools and techniques, specifically in relation to molecular marker applications in crop breeding, and the potential and prospects of bioinformatics in genomics.

2. Analysis of Genetic Diversity using Multiple Datasets

Ascertaining genetic diversity, particularly among germplasm accessions and elite breeding materials, is an important aspect in crop breeding. Accurate assessment of the levels and patterns of genetic diversity can be invaluable for: (i) identifying diverse parental combinations to create segregating progenies with maximum genetic variability for further selection; and (ii) introgressing desirable genes from diverse germplasm into the available genetic base. Significant emphasis is being paid towards comprehensive analysis of genetic diversity in numerous crops, including major field crops such as wheat, rice, maize, barley, soybean etc. Diverse data sets have been employed to analyze genetic diversity in crop plants; most important are: pedigree data, morphological data, biochemical data (mainly isozyme analysis), and more recently, DNA-based marker data. These data sets provide different types of information, and have their own strengths and limitations.

Multivariate analytical techniques, that simultaneously analyze multiple measurements on each individual under investigation, are widely used in analysis of genetic diversity irrespective of the data set. Among these, cluster analysis, principal component analysis, principal coordinate analysis, and multidimensional scaling are, at present, most commonly employed. There are broadly two types of clustering methods: (i) Distance-based methods, in which a pair-wise distance matrix is used as an input for analysis by a
specific clustering algorithm, leading to a graphical representation (dendrogram) in which clusters may be visually identified; and (ii) Model-based methods, in which observations from each cluster are assumed to be random draws from some parametric model, and inferences about parameters corresponding to each cluster and cluster membership of each individual are carried out jointly using standard statistical methods such as maximum-likelihood or Bayesian methods. Despite the constraints imposed by influence of specific distance measure, clustering algorithm and/or graphical representation chosen, distance-based methods are still widely preferred probably because they are relatively easy to apply for analysis of various data sets.

A number of publications are available with regard to the use of multivariate analysis for genetic diversity studies in crop plants. There are still some important issues to be addressed regarding: (i) Judicious and effective use of different types of variables like continuous, discrete, ordinal, multi-state, binomial etc. in genetic diversity studies; (ii) Proper choice of a genetic distance measure, clustering algorithm(s) or multivariate statistical technique(s) in analysis of data; and Objective determination of genetic associations/clusters of individuals. Strategies required to address the above issues might vary depending on the genetic materials being analyzed and the objectives of the experiment. During my presentation, I shall discuss some practical issues related to the above specific issues, and the possible ways to increase the objectivity in data interpretation, particularly in relation to application of molecular marker data for analysis of genetic diversity (Mohammadi and Prasanna, 2003).

3. QTL Analysis – Bridging Quantitative Genetics with Molecular Biology

Two significant developments during the 1980s have caused a paradigm shift in the study of quantitative traits in plants and animals. The first was the discovery of extensive, yet easily visualized, molecular polymorphism using DNA-based markers. The second was the advent of a QTL (Quantitative Trait Loci) analysis, that allows associations between the quantitative trait and marker alleles segregating in the population using molecular marker technology and genetic linkage maps. Initially, RFLPs were used as markers, and these were followed by PCR-based markers such as microsatellites and AFLPs. QTL analysis has been carried out in a range of crop species for traits related to yield, quality, disease and insect resistance, abiotic stress tolerance etc. These analyses are already providing important clues to explore the number, position and effects of QTLs influencing agronomically important quantitative traits in both plants and farm animals (Tanskley, 1993). Putative locations and DNA markers closely linked to QTLs have opened up the possibilities for isolation and characterization of QTLs, introgression of QTLs into breeding lines or germplasm, and marker-assisted selection for QTLs in breeding. Although success stories in relation to effective utilization of QTL information in breeding programmes, at present, are extremely limited, if not negligible, it must be kept in mind that this specific area of molecular breeding is hardly two decades old.

QTL analysis is essentially supported by three ‘pillars’: (i) molecular marker technology that enables genotyping; (ii) scoring of the trait of interest (phenotyping); and (iii) statistical methodologies that facilitate analysis of association of the trait with markers. In recent years, statistical methods using interval mapping (based on the joint frequencies of
adjacent markers and a putative QTL flanked by two or multiple markers) have enhanced the power of QTL mapping. The data can be analyzed by statistical methods such as maximum likelihood, non-linear regression or a combination of maximum likelihood and multiple regression called composite interval mapping (Jansen and Stam, 1994; Zeng, 1994). Several software packages, such as Mapmaker/QTL, MAPQTL, QGENE and QTL Cartographer, are now available for QTL analysis. For most QTL mapping methods, analyses can be implemented in common statistical software packages, such as SAS. However, when the number of markers is large, specialized software packages, as mentioned above, may be needed.

Despite significant advances, there are still important problems to overcome in methods of QTL mapping, such as mistakes and misinterpretations owing to too much confidence in QTL position, inadequate significance levels and use of over-simplified models (Kearsey and Farquhar, 1998). New approaches are being explored by researchers to increase the resolution of QTL mapping through selection of valid mapping populations, using improved statistical packages and employing a multi-step strategy. The next major challenge is how to effectively link the growing information on QTLs with molecular biology. This would require map-based cloning of QTLs, and possible identification of candidate genes controlling specific quantitative traits by making use of information either already available or being generated on mapped genes, mutants and ESTs (Expressed Sequence Tags).

4. Computational Applications in Molecular Breeding
Besides linkage mapping and QTL analysis, statistical methodologies and computer simulations have a powerful role to play in molecular breeding, including (i) utilization of molecular marker data of the parental lines in prediction of single-cross hybrid performance; (ii) molecular marker-assisted selection (MAS), particularly for complex traits, in breeding programmes. Several studies have been carried out in recent years analyzing the association between molecular marker divergence and hybrid performance. Although divergent views have been expressed about the utility of such an analysis in predicting hybrid performance, evidences are indeed available to suggest that precise evaluation of genotypic differences using molecular markers may be useful for preliminary selection of loci and alleles for possible improvement of hybrids (Stuber et al., 1999). Statistical methodologies such as Best Linear Unbiased Prediction (BLUP) offer promise in improving the predictive ability of parental marker data in relation to hybrid performance (Bernardo, 1994).

Theoretical and analytical investigations have shown that maximum rate of improvement with respect to quantitative traits may be obtained by integrating both phenotypic and marker data. The potential selection efficiency in such a strategy depends on the heritability of the trait, the proportion of additive genetic variance associated with the marker loci, and the selection scheme (Lande and Thompson, 1990). Recent studies have also indicated statistical limitations on the efficiency of marker-assisted selection, which include the precision of the estimated associations between marker loci and QTLs as well as sampling errors in estimating weighting coefficients in the selection indices. Knapp (1998) developed the theory for estimating the probability of selecting one or more
superior genotypes by using MAS and included a parameter to estimate the cost efficiency of MAS relative to phenotypic selection. He reported that a breeder using only phenotypic selection must test 1.0 and 16.7 times more progeny than a breeder using MAS to be assured of selecting one or more superior genotypes. Thus, MAS can substantially decrease the resources needed to accomplish a selection goal for a low to moderate heritability trait when the selection goal and selection intensity are high.

5. Processing and Utilization of Genomic Data

‘Genomics’ seeks to understand how genes and genomes are structured, how they function, and how they have evolved. Genomic research is profoundly altering the way biologists think about living things. At this time, this branch of science is driven primarily by the human genome project and its spin-offs (the sequencing of entire microbial genomes and other ‘model’ organisms such as *Drosophila*, *Caenorhabditis* and *Arabidopsis*). Enormous progress has been made in automating the identification of genes in genomic sequences. However, building accurate models of genes from the sequences still requires a lot of human, ‘hands-on’ effort. There are two general approaches to gene finding: (i) homology-based approach that includes the use of known mRNA sequences as well as gene families and inter-specific sequence comparisons; and (ii) *ab initio* approach that includes detection of exons and other sequence signals, like splice sites, by various computational methods within the sequence being analyzed. Approaches that use a combination of statistical and heuristic methods to recognize genes and gene features are prevalent; hidden Markov models, neural nets, and Bayesian networks are among the methods used. Despite some successes in gene annotation using these approaches, the methods are not completely fool-proof due to complexities associated with biological systems coupled with inherent limitations of experimental/computational tools and techniques (errors in sequencing, statistical biases etc.).

Genome sequencing and functional genomics projects in a wide array of organisms are making available new information resources. These are in a wide range of formats that allow the data to be browsed or downloaded, but not necessarily analyzed effectively in conjunction with other information sources. Conceptual modeling of genomic information is another active area in which statistical and computational tools and techniques have a significant role to play. International efforts are currently under way to develop a Genome Information Management System (GIMS), which can be used as a scientific data warehouse.


To deal with the ‘tidal wave of data’ in biology, a new discipline has emerged in recent years, popularly known as ‘bioinformatics’. Bioinformatics is the computer-assisted data management discipline that helps to gather, analyse and represent this information in order to educate ourselves, understand life’s processes in the healthy and disease states, and find new or better products for the benefit of mankind. Bioinformatics represents the convergence of two technology revolutions: the explosive growth in biotechnology, paralleled by an equally explosive growth in information technology. We already have complete genome sequence information for several microorganisms, and similar
information shall be soon available for a range of plants and animals. The challenge is to ‘capture’ relevant biological information from the sea of data and make it readily accessible. In this context, bioinformatics has a vital role to play in: (i) Global and local sequence alignments in genome projects, using algorithms such as Needleman-Wunsch and Smith-Waterman; (ii) Finding genes through computational methods; (iii) Unraveling gene functions through multiple sequence alignments and searches; (iv) Classification of putative proteins and functional assignment; (v) Phylogenetic analysis and comparative genomics; and (vi) Developing database tools for biological data mining.

There could be several exciting applications stemming from bioinformatics in the coming decades. These shall include:

- Functional bioinformatics, which refers to the development of ontologies or concept classifications that are propelled by computational algorithms based on functions of biomolecules as their inputs and outputs.
- Integration of statistical genetic methods, sequence information, gene variability in populations, with epidemiological data.
- Modeling of genetic and metabolic networks in plants and animals for designing accurate and predictive computer-based models of biological function.

The ultimate objective is not just to obtain exciting insights into the structure, nature and dynamics of genomes of diverse organisms (which is by itself a tremendously exciting and challenging goal!), but also to effectively address various problems encountered in medicine and agriculture. This cannot be possibly achieved without reliance on relevant software and database systems to design gene arrays, track materials, collect and analyze, and interpret data from gene expression studies. Such systems should have to catalog the expression behavior of probably hundreds of genes in a single experiment and, subsequently, make comparisons across tissues, developmental and pathological states, or cellular perturbations. Although this particular field is still in its infancy, there are exciting possibilities for carrying out biological experiments in silico.

7. Concluding Remarks

An increasing number of databases and feature-packed software packages are now available to deal with a range of applications in genetics and biotechnology, including analysis of genetic diversity using molecular marker data, QTL mapping, DNA sequence analysis, protein function prediction etc. The success of biotechnology would undoubtedly depend a great deal on these software packages, databases, and networking systems. Comprehensive computational tools will be needed to integrate information regarding genotypic performance, pedigree relationships, and germplasm diversity so that genomic data can be interpreted in ways that are useful to agricultural scientists. We must also keep in mind that statistical/computational packages, however powerful they might be, provide little, if anything to guide users when it comes to choosing the ‘right tool’ for the job and the interpretations of possibly complex results often encountered in genetics and molecular biology. Choosing a proper method for data analysis, based on rational choice rather than by habit or based on its popularity, is extremely critical. Equally important is the next step, that is, proper interpretation of results. How to differentiate biologically meaningful information from artifacts? In other words, how to sieve the ‘grain’ from the ‘chaff’? The fact is there is no simple solution or a magic formula to
address the above question. Nor there is likely to be a consensus as to which statistical or computational tools or techniques are indeed most appropriate under a given situation. Nevertheless, considerable thrust is being given to finding software solutions and packages that are powerful, effective and user-friendly. Probably, what we still require and what we may also soon shall see would be well-designed ‘electronic assistants’ that can intelligently interact with a user to devise effective analytical strategies and that collaborate on the interpretation of analysis results. Such an intelligent advisory system might be able to monitor the user’s actions, suggest appropriate next steps, or alternate approaches, point out possible pitfalls and provide background information as required or requested. Although such a system can theoretically and practically provide context-sensitive help, it is important that the ‘control’ of using a right tool for a right job for generating meaningful information must ultimately remain with the user.

Concerted efforts are certainly required to make bioinformatics an active and integral component of our R&D efforts in genetics, plant breeding and biotechnology. India, undoubtedly, has the required talent in IT, However, extremely few scientists in the country can claim to have a strong background in both biology and computational science that is needed to formulate effective computational models and packages for providing solutions to complex biological problems. There is a dearth of ‘mentors’ who can possibly train the next generation of researchers with bioinformatics expertise. This important rate-limiting factor can possibly be addressed by enhancing the capacity of the country in relation to high-throughput experimental technologies, and making necessary institutional adjustments for faculty orientation, research and teaching skills in relation to genetics, biotechnology, statistical and computer sciences, and integration of bioinformatics as an integral component of our research and educational system. What is required is a human resource base that can perform two critical roles: (i) effectively applying the existing computational tools to achieve new insights about genetics and molecular biology; and (ii) developing new statistical and computational algorithms and databases for enhancing the precision and efficiency in biological experimentation, data generation and interpretation.

References