Symposium on
Statistical and Computational Genomics

Chairman   : **Prof. Prem Narain, Executive President, ISAS**
Conveners : **1. Dr. Rajender Parsad, IASRI, New Delhi**
              2. Dr. B.M. Prasanna, IARI, New Delhi
              3. Dr. V.K. Shahi, RAU, Pusa, Samastipur

Three papers covering various aspects related with the theme of the symposium were presented by the following speakers:

1. Statistical Genomics for Crop Improvement: Opportunities and Challenges – B.M. Prasanna

2. Design and Analysis for 2-colour Microarray Experiments – Rajender Parsad

3. Bioinformatics in Agriculture – V.K. Bhatia

After long deliberations, following recommendations were emerged out:

1. Computing environment (both in terms of software and hardware) should be strengthened at par with International level so as to enable the researchers to perform voluminous data analysis of available in Genomics.

2. Identify the faculty members with aptitude in statistical genomics and provide advanced training to them in leading Institutions.

3. To introduce M.Sc. programme in Bioinformatics with special reference to Mathematics, Statistics and Computational Biology.

4. To create infrastructure for National Repository of Agricultural Genomic databases.

5. Concerted efforts need to be made for development of efficient and robust experimental designs and efficient analytical technique for single and multi-factor microarray experiments.

6. Basic and Applied research in the areas of proteomics, functional genomics also needed to be undertaken through a strong network of inter-institutional programme.

ABSTRACTS OF THE PAPERS PRESENTED

1. Statistical Genomics for Crop Improvement: Opportunities and Challenges

B.M. Prasanna

In the last one decade, exciting advances have been made in statistical genomics, besides development of high throughout genotyping. Powerful statistical methods and tools are immensely aiding research progress in molecular plant breeding and in bridging the genotype-phenotype divide. This is largely due to the evolution in statistical and computational means for analyzing the patterns of molecular diversity, formulating core collections, understanding genotype $\times$ environment interactions, identifying QTLs, detecting epistatic interactions, undertaking association mapping, and analyzing microarray data. The presentation will review the national and international developments in these important areas, with particular focus on crop plants, and shall identify critical gaps for strengthening statistical genomics in the Indian context.

The new focus on genomics has also highlighted a particular challenge: how to integrate the different views of the genome that are provided by various types of experimental data and provide a proper biological

perspective that can lead to crop improvement. Mapping and studying the genetic architecture of complex traits, and understanding the dynamic network of gene interactions that determine the physiology of an individual organism over time is another major challenge that requires novel, quantitative and testable statistical solutions.

Besides development of comprehensive computational tools to integrate information regarding genotypic performance, pedigree relationships, germplasm diversity and genomic data, there is also an immense need to develop a new "breed" of geneticists and statisticians in developing and implementing novel, more effective and efficient translational bioinformatic systems that the scientists can routinely use in breeding strategies. The availability of efficient computational algorithms/softwares is essential to the scientific community. However, it is equally important that these tools are applied with thorough understanding of the genetic data and the tools themselves.

Indian Agricultural Research Institute, New Delhi

2. Design and Analysis for 2-colour Microarray Experiments

Rajender Parsad and V.K. Gupta

Microarrays are microscopic arrays of single-stranded DNA molecules immobilized on a solid surface by biochemical synthesis. **DNA microarrays** are created in two basic forms viz. (i) by DNA depositions and (ii) by in situ synthesis of oligonucleotide arrays. These are also known as *DNA chips*, *gene chips*, *biochips*, *DNA microarrays* or simply the *arrays*. Microarray is an important genomics tool that can identify the expression of several thousand genes at a time.

In 2-colour cDNA microarray experiments, four basic experimental factors viz., array (A), dye (D), variety (V) and gene (G) are studied. These four factors give rise to 15 effects that include 4 main effects, 6 two-factor interactions, 4 three-factor interactions and one four factor interaction. But all the main effects and selected two-factor interactions viz. array-gene interaction (AG), dye-gene interaction (DG), variety-gene interaction (VG) are the seven effects of interest to the experimenter. An important and common question in DNA microarray experiments is the identification of differentially expressed genes, that is, genes whose expression levels are associated with a

response or covariate of interest. Designing of microarray experiments is an important issue to get precise comparisons of variety ´ gene interactions. Reference designs, alternating loop designs and dye-swap designs are the most commonly used designs for these experiments.

The investigations carried out so far deal with the situations where same set of genes is spotted on each array in microarray experiments. Therefore, genes/ gene specific effects (G, AG, DG, VG) are orthogonal to global effects (A, D, V). Therefore, optimality aspects of designs for microarray experiments have been studied by taking only array, dye and variety effects and leaving gene specific effects from the model. Designs that are efficient under the model containing only global effects are also efficient under the model containing both global and gene specific effects. Efficient designs have been obtained for these experiments under the assumption that the dye effects are orthogonal with respect to variety effects under a fixed effects model. Catalogue of efficient block designs for microarray experiments are now available in the literature. Further, array effects may be random, and as a consequence, model becomes a linear mixed effects model. The robustness aspects of efficient designs obtained under a fixed effects model have been investigated under mixed effects model. There is need to obtain designs by considering the dye effects also in the model.

In the research investigations carried under a mixed effects model, the ratio of inter and intra block variances has been assumed as constant. One may consider putting a beta prior on the ratio of the variance components and then derive efficient designs after taking the expectation with respect to beta distribution. This may be achieved through Bayesian inference.

Further, in all the investigations carried so far, it has tacitly been assumed that the variability in gene expression is constant across all the genes. Depending upon the underlying biology, the gene expressions may be heteroscedastic and depend upon the gene of interest. This is another important issue which needs attention. This amounts to obtaining efficient block designs under a heteroscedatic set up.

In most of the investigations carried out so far it has been assumed that all the genes are spotted on all the arrays. There is a need to obtain efficient designs when genes spotted on each array differ, i.e. array and gene effects are non-orthogonal.

The above discussion is for the experimental situations where there is only one factor that is causing variation in gene expression levels, i.e. variety (different types of tissues, drug treatments or time points of a biological process). There, however, do occur experimental situations, where it is desired to compare more than one factor for studying the gene expression levels. For example, one may be interested to study and compare the two mutants at times zero hour and 24 hours. The interest is in measuring the changes over time. Therefore, there are two factors viz. varieties (two mutants) and time (0 hour and 24 hours) and there is a need to obtain efficient designs for factorial and time-course microarray experiments.

For factorial microarray experiments, balanced factorial designs may be useful. The application of designs for balanced factorial experiments in 2-colour microarray experiments require that these designs should be constructed with block size two since only two varieties/ treatment combinations can be accommodated on one array. Some methods of construction of designs for balanced factorial experiments with block size two are available in the literature. These methods are heuristic in nature and give only designs for very few parametric combinations. It is, therefore, required to obtain efficient designs for balanced factorial experiments with block size two that can have an application in factorial or time-course microarray experiments. The efficient designs for balanced factorial experiments are only useful for the situations where a natural baseline does not exist for at least one of the factors, like gender lacks a natural baseline. This is known as orthogonal parameterization problem. For orthogonal parameterization, EGD designs may be useful and catalogues of EGD designs with block size two need to be prepared.

There do exists situations, where a natural baseline or null state exists such as the situation involving two mutants where one proliferates a particular disease and other does not. The mutant that does not proliferate into disease is baseline. In toxicological study with binary factors, each representing presence or absence of a toxin factor, absence can be regarded as a natural baseline level of each factor. Null state or baseline level of a factor need not strictly mean zero level on some scale, but may as well refer to a standard or control level like the one currently being used in practice. This is known as baseline parameterization. Definitions of main effects under the two parameterizations (orthogonal and baseline) are entirely different. Therefore, the designs that are efficient for one parameterization may not be optimal/ efficient under other parameterization. Problem of obtaining efficient designs under baseline parameterization has received a little attention. There is a need to put research emphasis on obtaining efficient designs for factorial microarray experiments for baseline parameterization.

It is also required to prepare a catalogue of such efficient designs to serve as a ready recknoer for the experimenters. Further, it is also required to arrange these designs in row-column structure so as to display both the variability factors viz. arrays and dyes.

Several other statistical issues involved in the analysis of gene expression data include data quality, data analysis and validation. The biological question of differential expression can be stated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. Researchers are interested in determining the direction of rejection of the null hypothesis, which is, in determining whether genes are over or under expressed. Generalized p-values are being used for these situations. One may also be interested in testing several hypotheses simultaneously for each gene. Another aspect in which one may be interested is the two-dimensional testing problem where several hypotheses are tested simultaneously for each of thousands of genes. This problem of multiple hypothesis testing requires attention of the statisticians. The genes are spotted on each array one along the other. Therefore, the fluorescence intensities from the nearby points may be correlated. Therefore, developing analytical techniques of data in the presence of spatially correlated observations is an unprecedented challenge. Another important problem in microarray experiments is classification of genes using the data from microarray experiments. Keeping in view the above, efforts will be made to obtain efficient designs for microarray experiments and to develop statistical analysis procedures for identification of differentially expressed genes.

Some other issues are how many biological samples should be taken and to what extent does biological averaging holds? Further, with the availability of multiple platforms employing cDNA or oligonucleotides, which may also differ in probe preparation methods and array surface chemistry, raises the question of cross-platform agreement in gene

expression measurements. Several studies have been carried out for comparing different microarray platforms.

Minimum Information about a Microarray Experiment (MIAME) is needed to enable interpretation of results of experiment unambiguously and to reproduce experiment. Six most critical elements contributing towards MIAME are: (i) Raw data for each hybridization, (ii) Final processed (normalised) data for the set of hybridisations in the experiment, (iii) Essential sample annotation including experimental factors and their values, (iv) Essential sample annotation including experimental factors and their values, (v) Sufficient annotation of the array and (vi) Essential laboratory and data processing protocols. The public repositories Array Express at the EBI (UK), GEO at NCBI (US) and CIBEX at DDBJ (Japan) are designed to accept, hold and distribute MIAME compliant microarray data.

Several statistical packages are now available that deals with microarray experiments. To name a few: JMP Genomics, SAS Genomics, SAS Macros AnovArray, GenStat 12, S+ArrayAnalyzer , TM4 suite, Bioconductor, etc. For E-leraning on *Statistical Genomics*, a new link has been created on Design Resources Server (ww.iasri.res.in/design) that aims to provide support towards statistical and computational aspects of plant genomics. The main purpose of initiating a link on Statistical Genomics is to provide an E-learning platform to the experimenters in their analysis of data.

Indian Agricultural Statistics Research Institute, New Delhi

3.   Bioinformatics in Agriculture

V.K. Bhatia

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an exponential growth in the biological data generated by the scientific community. The appropriate analysis and management of genomic data has resulted very useful information and knowledge about the underlying phenomenon. With the advent of newer informatics techniques this needs to be further looked into for extracting additional deeper knowledge. Thus, the information generated from biological experiments is essentially required to be stored, organized, understood and analyzed for furthering the knowledge and insight of the genomic

data. This search of newer knowledge has led to the emergence of an integrated discipline, called Bioinformatics. This rapid technological advancement has impacted agriculture significantly and thus, the research community involved in agricultural sciences need to participate and contribute to the ensuing global bioinformatics revolution.

The application of bioinformatics has been carried out in various ICAR labs for agricultural/biological sciences. Looking at the current status of bioinformatics research at International and National levels some research gaps pertaining to agriculture have been identified. The gaps basically identified are related to computational biology, statistical genomics and development of newer algorithms for handling voluminous genomic and proteomic data. In order to fill up these gaps ICAR initiated a process of creating a Centre of Bioinformatics for processing genomic data collected at ICAR institutes across all species. The present article deals with the importance of bioinformatics in agriculture and the initiatives taken along with identifying the future research strategies. The emphasis for development of a National Agricultural Bioinformatics Facility under ICAR to meet the challenges ahead in the field of bioinformatics in agriculture has been highlighted.

The bioinformatics initiative will enable the flow of information from the partners in terms of database and problems to be core group for eliciting meaningful and useful inferences by leveraging the tools of information technology for arriving at solutions to their problems. This will not only expedite layout of experiments and generating more such data on a continuous and ongoing basis but also serve as a veritable knowledge base. The partners would be in a position to interact with one another directly or indirectly for solving problems of mutual and common nature. The information so gathered will be warehoused in the centre. The centre will also take care of the needs for collaboration with other Indian and international organizations for enhancing the database, by way of human resource development and also widening the perspectives of bioinformatics research by deeper insights into various aspects. Its collaboration is also expected to enable the country to emerge as a global player for genomic research in agriculture.

Indian Agricultural Statistics Research Institute, New Delhi

Symposium on
Statistical and Informatics Perspective of Climate Change

**Chairman : Prof. Bikas Sinha, Visiting Prof. ISI, Kolkata
Conveners : 1. Dr. P.K. Malhotra, IASRI, New Delhi
            2. Dr. R.C. Agrawal, NBPGR, New Delhi**

Three papers covering various aspects related with the theme of the symposium were presented by the following speakers:

1. Spatio-temporal data mining for monitoring of climate change – Anil Rai

2. Experimental designs for mitigation and adaption of climate change – Rajender Parsad

3. Modeling climate effects – V.K. Bhatia

After detailed discussions, the following recommendations emerged out:

 (i) Multivariate analysis and data reduction techniques need to be used for monitoring climate change.

 (ii) Data sets of spatio-temporal variables on climate as well as derived parameters based on spectral data from MODIS must be used in climate change studies.

(iii) For development of cultivars/agricultural practices for mitigation and adaption of climate change, study on designing of experiments and analysis of experimental data should be taken up vigorously.

(iv) Data modeling should be given emphasis for identifying the variables for climate change studies.

 (v) Development of knowledgebase required for climate modeling and climate impact assessment studies.

ABSTRACTS OF THE PAPERS PRESENTED

1. Spatio-Temporal Mining for Monitoring Climate Change

Anil Rai, K.K. Chaturvedi and P.K. Malhotra

Climate is usually defined as the "average weather", or more rigorously, as the statistical description of the weather in terms of mean and variability of relevant parameters over periods of several decades (typically three decades as defined by WMO). These parameters are most often surface variables such as temperature, precipitation and wind, but in a wider sense the "climate" is the description of the state of the climate system. The climate system consists of five major components as (a) the atmosphere, (b) the oceans, (c) the terrestrial and marine biospheres, (d) the cryosphere (sea ice, seasonal snow cover, mountain glaciers and continental scale ice sheets) and (e) the land surface. These components interact with each other, and through this collective interaction, determine the earth's surface climate. It has been observed that due to global warming, global average temperature is continuously increasing. There

is increase in global average temperature by one degree from 1961 to 1990. As a consequence of this sea level is increasing and overall area under snow cover is decreasing. Climate change scenarios for future can be best project using magic quadrant of Special Report on Emission Scenario (SRES). Agriculture is one of the most venerable sectors, which will have high impact of this global warming. Increase in CO_2 in the atmosphere will increase photosynthesis capacity of the plants, decreases stomatal conductance, enhanced water use efficiency, altered photosynthesis partitioning and finally its impact on plant depends on elevation and C3/C4 plants type. Increase in temperature will have negative impact on crop yield especially during winter season. In this process, there may be increase in yield, which will have positive effect on paddy crop but at the same time there will be increase in extreme events which will have negative impact on agricultural production. Further, aerosols in the atmosphere will increase and as a consequence of this there will be reduced solar radiations on the earth, which will have negative impact on agricultural production. Therefore, the combined effect of climate change will be very complex and difficult to quantify in relation to agriculture. However, monitoring of climate and its effect can help in better planning and minimizing the damage to global agriculture. Construction of Ocean Climatic Index (OCI) is one of the important techniques to monitor the climate and its impact at global level. OCI captures ocean and land relationship. It is based on time series data and it summarises behaviour of selected area on land surface based on climatic parameters of ocean. El Nino effect is one of the well known pattern of changing direction of trade winds in such a way that it brings drought in Australia, warmer winters in North America, flooding in coastal Peru and increased rainfall in east Africa. Important climatic variables are Sea Level Pressure (SLP), Sea Surface Temperature (SST) and precipitation. One of the most important dependent variable, which indirectly measures the plant growth is Net Primary Production (NPP). NPP measures net assimilation of atmospheric Carbon Dioxide in to organic matter by plant and helps in understanding global carbon cycle at regional and global level. It is driven by solar radiation and constraint by precipitation and temperature. These variables are measured at spherical grid level. During pre-processing of this data, monthly Z-score are calculated to remove seasonality and spatial as well as temporal correlations. There are other numbers of known climatic indices such as Southern Oscillation Index (SOI), North Atlantic Oscillation (NAO) and Dipole Mode Index (DMI) etc. DMI is very important for prediction of rainfall during monsoon in India. Singular Value Decomposition (SVD) analysis is one of the important techniques for development of OCI. It is similar to Principle Component Analysis (PCA) and captures the spatial and temporal pattern in the climatic data. This is also known as Empirical Orthogonal Function (EOF). But, it also has number drawbacks. SVD is able to detect only few strongest patterns as strong patterns mask the detection of weaker patterns. Interpretability is poor due to orthogonal vectors. Further, analysis cannot take into accounts of different lags in spatial patterns. Recently, graph based clustering has been used to avoid these drawbacks. Most popular techniques are Shared Nearest Neighbour (SNN), Jarvis-Patrick Clustering etc. These cluster based OCI's are found to be very effective in monitoring of climate at global level. Now Moderate Resolution Imagine Spectro-radiometer (MODIS) provides wealth of data on various climatic parameters along with derived variables which can be used in development of OCI's for monitoring the global climate more effectively.

Indian Agricultural Statistics Research Institute, New Delhi

2. Experimental Designs for Mitigation and Adaptation Strategies of Climate Change

Rajender Parsad[1], V.K. Gupta[1] and
Raj. S. Malhotra[2]

Climate change refers to the changes in the mean and /or variance of the parameters such as temperature, rainfall, presence of greenhouse gases, etc. over a long period of time. Obviously, this change should be identifiable using statistical tests. Therefore, the first and foremost requirement of studying climate change is the data availability of given parameters at a given location over an extended period say at least 3 to 4 decades. To see the spatial pattern over geographic locations, the data availability should be at different locations geographically. In the study of climate changes, there are two approaches viz.

(i) to detect the changes over time and space and

(ii) to study the causes of those changes.

The first one can be achieved by studying the trends through saptio-temporal models. In this case, no reasons are to be assigned. According to Levine and Berliner (1999, *Journal of Climate*, 12(2), 564-574), the main difficulty in early detection of changes resulting from human induced/ anthropogenic forcing is that the natural variability overwhelms the climate change signal in the observed data. To overcome this problem, fingerprint procedures have been developed to express the climate data in terms of low-dimensional signal patterns. Fingerprint approach is a mathematical procedure for optimally detecting a climate change signal above the background natural climate variability noise. For detection, in statistical perspective, it is testing of hypothesis that the change is not due to natural variability. Some people also profess Bayesian approach to multi-model information processing for developing climate forecasts.

To study the causes of changes, one requires the data on the variables and one has to develop models for cause and effect relationships. For establishing these relationships controlled comparative experiments may be helpful.

The effect of global warming and climate change on agriculture would be in terms of, rise in global temperature, elevated levels of carbon dioxide in the atmosphere, altered rainfall patterns, greater severity and frequency of extreme weather events, including droughts and floods, etc. This will result in decrease in availability of water, may be more severe droughts, sudden weather fluctuations, increased vulnerability to diseases and insect pests, etc. There may be swings in temperature and rainfall. There may be heavy summer and winter rainfalls but with only small changes in annual rainfall in totality.

Therefore, the need is to strengthen the adaptive capacity of the communities, and resilience of farming systems, in short, mitigation and adaptation.

For deciding upon mitigation and adaptation measures is to know the effect of increased levels of CO_2, rise in temperatures, extended drought periods, we need to study the effects of these effects of climate change including swings on agriculture and food security. A lot of studies have been conducting for studying the effect of elevation in level of CO_2, rise in temperature, drought periods etc. taken one factor at a time. The interactions among CO_2, temperature and water can be substantial and the combined effect on the biological systems of several factors may not be predicted from experiments with one or a few factors. Therefore, it is required to conduct multi-factor experiments involving a large set of factors such as taking together different levels of elevation of CO_2, rise in temperature, drought periods etc. One such experiment was conducted with two levels of CO_2 (ambient and elevated level 510 ppm) two levels of temperature (ambient and elevated temperature of 2^0C) and two levels of droughts (present condition and induced drought in late spring summer) at Brandbjerg, Copenhagen, Denmark during 2005-2007 using split plot designs in 6 replications on a grassland eco-system. For details on this experiment a reference may be made to Mikkelsen *et al*. (2008, *Function and Ecology*, 22, 185-195). In this experiment only the extreme levels of different factors have been used and it does not provide any information on the intervening levels of these factors. Further, the effects of factors of climate change may vary from region to region and crop to crop. Therefore, region specific experiments should be conducted using varying levels of different factors of climate change. This requires efficient designing of experiments under scarce resources.

Once these effects are known, the mitigation strategy is to identify/ develop varieties which are drought tolerance, cold tolerant, resistant against diseases and pests, varieties which can sustain the high concentrations of CO_2, etc. In other words, we have to develop/ identify varieties that are resistant to biotic and abiotic stresses. We have also to think of identifying improved cropping systems and practices that make efficient use of natural resources. In all these efforts, statistical designing and analysis of experimental data may be of help. In the present talk, we shall concentrate on statistical issues related to design and analysis of experiments for identification of cultivars that are tolerant to biotic stress (living organisms which can

harm plants, such as viruses, fungi, and bacteria, and harmful insects, etc.) and/or abiotic stress {negative impact of non-living factors on the living organisms in a specific environment, for example, extreme temperatures (high and low) during whole or part of crop growth, drought, flood, and other natural disasters, soil conditions, etc.). In other words, crops/ cultivars containing genes that will enable them to withstand biotic and abiotic stresses need to be developed. Here, care should be taken that not only the cultivars are high yielding but also retain the nutrition quality in terms of protein, fat, carbohydrates, minerals contents. In this talk, we shall explain it through some situations that we encountered during our advisory services with the subject matter specialists.

[1]*Indian Agricultural Statistics Research Institute, New Delhi*

[2]*ICARDA, Aleppo, Syria*

3.   Modeling Climate Effects

V.K. Bhatia

Field crops are very much influenced by climate effects. It is more for perennial crops because they remain in the field for many years and are subjected to climate changes over a longer period of time. For a researcher, the primary objective is to assess the potential impacts of climate change and identify possible adaptation strategies. Thus, the main challenges in front of researchers are, how to study the effects of climate particularly for the situations of large number of variables related to climate parameters in comparison to very small number of observations of main character of interest of crop production or productivity. This problem, therefore, finally leads to the very important statistical problem of handling the situation of large p (number of independent factors) with small n (number of observations of the response variable). The solution of this problem has been studied by number of research workers. The contribution of these researchers has been reviewed in brief.

Mainly there are two issues, firstly that of model selection which deals with estimating the performance of different models in order to choose the best model and secondly having chosen a final model, estimating its prediction error on the new data. With these issues in mind, the applicability of different methodologies of Least Square Techniques (Regression), Lasso (Least Absolute Shinkage and Selection Operator) as given by Tibshirani (1996); Lars (Least Angle Regression) as given by Efron *et al*. (2004); Dantzig Selector by Efron (2007) and Dasso (Connections between the Dantzig selector and Lasso) by James *et al*. (2009) have been highlighted. The concept of the regression tree modeling has also been highlighted. In predictive modeling having large number of predictors, problem is to find "what subset of the effects which provides the best model for the data". In this area, the problems of identifying the best model are also brought to the notice of researchers. Finally, it is concluded that model selection is very important for building a high performance model and there is a need to develop methods for special data sets.

REFERENCES

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, (with discussion). *Ann. Statist*., **32,** 407-451.

Efron, Bradley, Hastie, T and Tibshirani, R. (2007). Discussion of "the Dantzig selector". *Ann. Statist*., **35,** 2358-2364.

James, G.M., Radchenko, P. and Jinchi, Lv. (2009). DASSO: connections between the Dantzig selector and Lasso. *J. Roy. Statist. Soc*., **B71**, 127-142.

Tibshirani, R. (1996). Regression shrinkage and selection via. the lasso. *J. Roy. Statist. Soc*., **B58**, 267-288.

Indian Agricultural Statistics Research Institute, New Delhi