A Data Mining Strategy for Exploring Cotton Genome – An Integrated Approach to Gene Prediction

M. M. Kshirsagar, Gurmit Singh and G. Balsubramani

Abstract –**This paper presents an integrated approach towards solving the problem of Gene Prediction. The Gene Prediction problem solving undergoes well defined stages starting with a DNA sequence as input and lab treatments and computational analysis go hands in hands throughout the process. Many bioinformatics tools are available for analysis at different stages of Gene Prediction, but a simplified and integrated approach is needed to support and speed up the task of a Life Scientist. A data mining strategy has been proposed in this paper to explore the comparatively less expressed Cotton Genome. The work is being carried out in CICR, Nagpur and Comparative Genomics is being used to predict Cotton Genes. Arabidopsis thaliana is being used as the reference plant. This strategy reveals the fundamentals and mathematics of the entire process based on which a complete software can be developed which can help in automating the process of Gene Prediction.**

Index Terms -- **Codon, DNA, Genome, Exon, Intron, mRNA, Protein, Splicing, Transcription, Translation**

I. INTRODUCTION

THIS paper presents a data mining strategy for exploring cotton genome while using an integrated approach to solve the problem of gene prediction. Cotton genome has not been fully explored as yet. Research is being carried out at different centers of cotton research in India and abroad. The process of gene prediction is a very much complex, tedious and a lengthy process. With the advent of Bioinformatics, the scientists involved in the research could have a sigh of relief, since different bioinformatics tools related to gene analysis and prediction come to the rescue of the scientists time to time. But this does not solve the problem. Basically, the entire gene prediction process has to be a well conceived blend of laboratory work and computational analysis so as to support the scientists and speed up their research by providing an integrated and efficient data mining facility. This paper is an attempt to reveal the fundamentals and mathematics of the entire gene prediction process and propose a data mining strategy based on which complete data mining software can be developed which can help in automating the process of gene prediction.

A. About Arabidopsis thaliana

Arabidopsis thaliana has been universally recognized as a model plant for study. It is a small flowering plant that belongs to the *Brassica* family, which includes species such as broccoli, cauliflower, cabbage, and radish. Although it is a non-commercial plant, it is favored among basic scientists because it develops, reproduces, and responds to stress and disease in much the same way as many crop plants. Scientists expect that systematic studies of Arabidopsis will offer important advantages for basic research in genetics and molecular biology and will illuminate numerous features of plant biology, including those of significant value to agriculture, energy, environment, and human health. Because of several reasons Arabidopsis has become the organism of choice for basic studies of the molecular genetics of flowering plants.

Arabidopsis thaliana has a small genome (125 Mb total), which already has been sequenced in the year 2000, [www.bioinformatics.nl], and it lacks the repeated, less-informative DNA sequences that complicate genome analysis. It has extensive genetic and physical maps of all 5 chromosomes (Map Viewer); a rapid life cycle (about 6 weeks from germination to mature seed); prolific seed production and easy cultivation in restricted space;

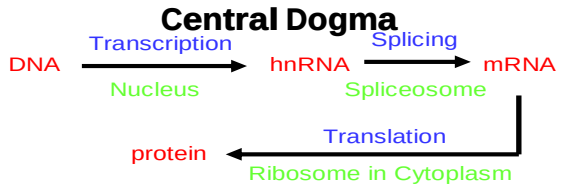
II. GENE PREDICTION PROBLEM

- Gene: A sequence of nucleotides coding for protein
- Gene Prediction Problem: Determine the beginning and end positions of genes in a genome

Mrs. M.M. Kshirsagar is an Asstt. Professor with the Department of CIT, Yeshwantrao Chavan College of Engineering, Nagpur 441 110, India (e-mail : manali_kshirsagar@yahoo.com).

Col..Gurmit Singh, Prof. and Head with the Department of IT and Electronics, AAIDU, Allahabad, India

Dr. G. Balsubramani is a Sr. Scientist with Central Institute of Cotton Research, Nagpur, India (e-mail: bala_amu@hotmail.com)



- Base Pairing Rule: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- Note: Some mRNA stays as RNA (ie tRNA, rRNA).

A gene is expressed in two steps : (Refer Fig 1)

1. Transcription: RNA synthesis
2. Translation: Protein synthesis

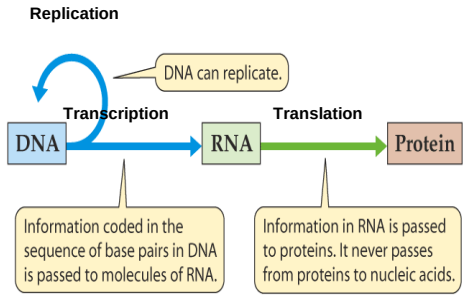


Fig. 1. DNA, RNA and the flow of Information

A. DNA → RNA: Transcription

DNA gets transcribed by a protein known as *RNA-polymerase.* This process builds a chain of bases that will become mRNA. RNA and DNA are similar, except that RNA is single stranded and thus less stable than DNA. (Refer Fig.2)

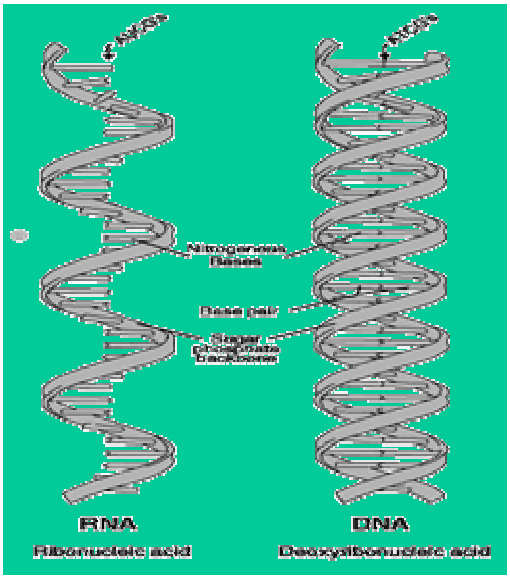


Fig. 2. DNA Vs. RNA

1) Terminology for Splicing

- Exon: A portion of the gene that appears in both the primary and the mature mRNA transcripts.
- Intron: A portion of the gene that is transcribed but excised prior to translation.
- Lariat structure: The structure that an intron in mRNA takes during excision/splicing.
- Spliceosome: An organelle that carries out the splicing reactions whereby the pre-mRNA is converted to a mature mRNA.

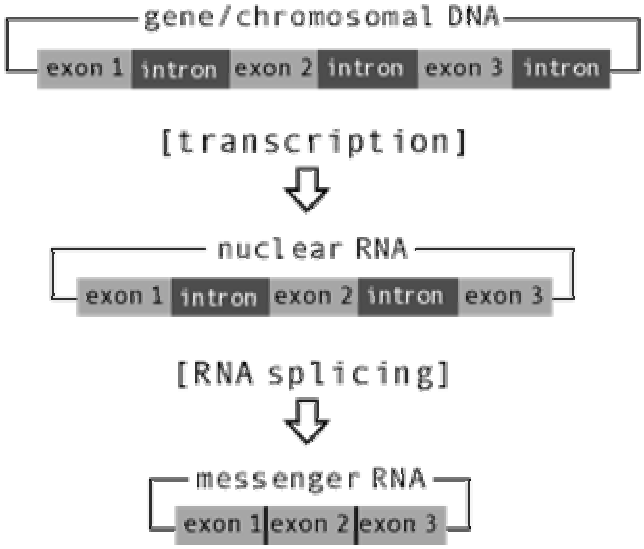


Fig.3. Transcription Process

- Unprocessed RNA is composed of Introns and Extrons. Introns are removed before the rest is expressed and converted to protein. (Refer Fig.3)
- Sometimes alternate splicing can create different valid proteins.
- A typical Eukaryotic gene has 4-20 introns. Locating them by analytical means is not easy.

2)Terminology for Ribosome

- Codon: The sequence of 3 nucleotides in DNA/RNA that encodes for a specific amino acid.
- mRNA (messenger RNA): A ribonucleic acid whose sequence is complementary to that of a protein-coding gene in DNA.
- Ribosome: The organelle that synthesizes polypeptides under the direction of mRNA
- rRNA (ribosomal RNA):The RNA molecules that constitute the bulk of the ribosome and provides structural scaffolding for the ribosome and catalyzes peptide bond formation.
- tRNA (transfer RNA): The small L-shaped RNAs that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA.

3)mRNA → Ribosome

- mRNA leaves the nucleus via nuclear pores.
- Ribosome has 3 binding sites for tRNAs:
  - A-site: position that aminoacyl-tRNA molecule binds to vacant site
  - P-site: site where the new peptide bond is formed.
  - E-site: the exit site
- Two subunits join together on a mRNA molecule near the 5' end.
- The ribosome will read the codons until AUG is reached and then the initiator tRNA binds to the P-site of the ribosome.
- Stop codons have tRNA that recognize a signal to stop translation. Release factors bind to the ribosome which cause the peptidyl transferase to catalyze the addition of water to free the molecule and releases the polypeptide.

4) Terminology for tRNA and proteins

- Anticodon: The sequence of 3 nucleotides in tRNA that recognizes an mRNA codon through complementary base pairing.
- C-terminal: The end of the protein with the free COOH.
- N-terminal: The end of the protein with the free NH3.
- The proper tRNA is chosen by having the corresponding anticodon for the mRNA's codon.
- The tRNA then transfers its aminoacyl group to the growing peptide chain.
- For example, the tRNA with the anticodon UAC corresponds with the codon AUG and attaches methionine amino acid onto the peptide chain.
- mRNA is translated in 5' to 3' direction and the from N-terminal to C-terminus of the polypeptide.
- Elongation process (assuming polypeptide already began)

tRNA with the next amino acid in the chain binds to the A-site by forming base pairs with the codon from mRNA

Carboxyl end of the protein is released from the tRNA at the Psite and joined to the free amino group from the amino acid attached to the tRNA at the A-site; new peptide bond formed catalyzed by peptide transferase.

Conformational changes occur which shift the two tRNAs into the E-site and the P-site from the P-site and A-site respectively. The mRNA also shifts 3 nucleotides over to reveal the next codon.

The tRNA in the E-site is released

GTP hydrolysis provides the energy to drive this reaction.

5) The Central Dogma

In going from DNA to proteins, there is an intermediate step where mRNA is made from DNA, which then makes protein.

DNA is kept in the nucleus, while protein sythesis happens in the cytoplasm, with the help of ribosomes. (Refer Fig.4)

6) RNA → Protein: Translation

Ribosomes and *transfer-RNAs* (tRNA) run along the length of the newly synthesized mRNA, decoding one codon at a time to build a growing chain of amino acids ("peptide")

The tRNAs have anti-codons, which complimentarily match the codons of mRNA to know what protein gets added next

But first, in eukaryotes, a phenomenon called splicing occurs in which introns, the non-protein coding regions of the mRNA are excised bringing together the exons, the coding regions so that functional, valid protein can form.

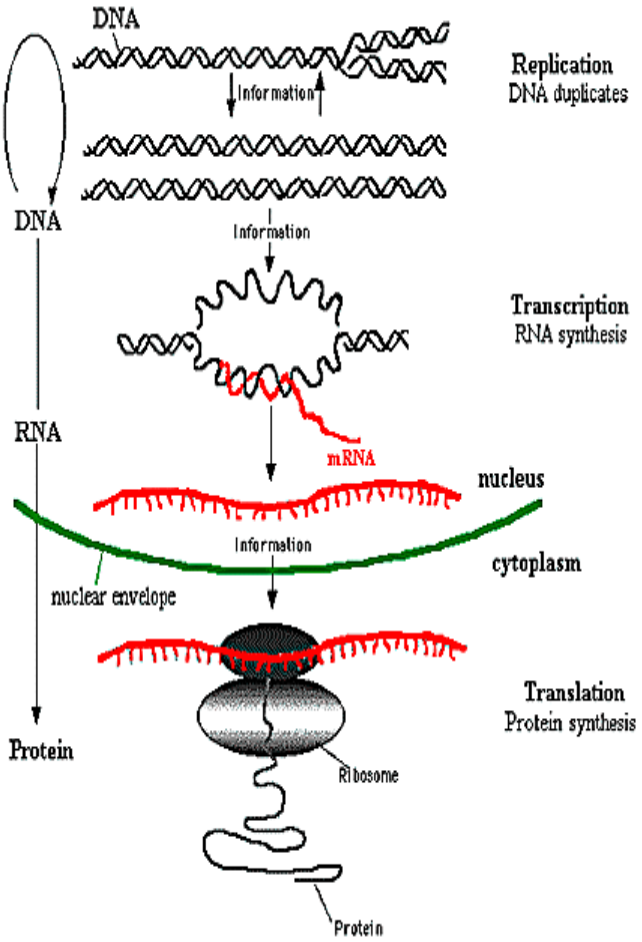


Fig. 4. The central dogma of molecular biology

III. INTEGRATED APPROACH TO SOLVE THE GENE PREDICTION PROBLEM

After understanding the gene prediction problem, the following generic steps can be taken to solve it taking DNA sequences as the input.

1. BLAST search – Identify similarity and dissimilarity between different sequences already available.
2. GC Content determination
3. Conversion from DNA to mRNA sequence
4. Identify restriction sites with various restriction endonuclease enzymes
5. Identify Introns and Exons
6. Identify Consensus Sequences
7. Identify promoter/terminal sequences
8. ORF identification
9. DNA to protein sequence conversion

While performing these steps/operations both the biological processes involved throughout the gene prediction process i.e. the laboratory treatments applied on the input DNA sequence and the underlying mathematics needs to be taken into consideration.

IV. DATA MINING STRATEGY FOR EXPLORING COTTON GENOME

As mentioned in the introduction, Cotton Genome has not been explored fully as yet. The work is in progress at different cotton research centers across the world, Central Institute of Cotton Research, Nagpur is one of them. The scientists use Arabidopsis plant as the model plant for gene prediction.

At CICR research is continued on Gene Isolation.

The genes on which work is in progress are:

Dehydrin (Dhn) – 3-5 DNA sequences

Osmotin (Osm) – 2-3 DNA sequences

Protease Inhibitor (PI) – 2 DNA sequences

Marker Gene

Ath A

Ath B

One DNA sequence is @400-2700 base pairs.

One base pair is A-T or G-C

1000 base pairs = 1kb

Step 1 : BLAST search – Identify similarity and dissimilarity between different sequences already available.

Due to the availability of Arabidopsis plant genome in the fully explored state, the data mining algorithm can provide an interface to BLAST tool for searching sequence similarity between the genes under isolation with that of Arabidopsis.

Step 2 : GC content determination

This step helps in identifying the potential gene sequences. In case of Eukaryotes, the amount of GC content is more in the genes. Hence, from the GC content, the potential sequences can be exposed to further analysis and non-potential sequences can be discarded.

This becomes a pattern matching module in the data mining algorithm i.e. identifying the GC content in the sequence.

Step 3: Conversion from DNA to mRNA sequence

As per the central dogma, DNA → Protein is not possible, hence DNA→ mRNA is the preprocessing step. The data mining algorithm will have a module for conversion of DNA sequence to mRNA string . This is based on a simple replacement concept of base T(hymine) with base U(racil).

Step 4: Identify restriction sites with various restriction endonuclease enzymes

Here, it is essential to know what restriction enzymes are.
Restriction Enzymes were discovered in the early 1970's and used as a defense mechanism by bacteria to break down the DNA of attacking viruses. They cut the DNA into small fragments while allowing the DNA sequence to be in a more manageable bite-size pieces.
It is interesting to note that restriction enzymes cut the DNA sequence at specific sequences/sites only. Thousands of such sites exist in a DNA sequence.

Some more interesting facts about restriction enzymes :

Type I restriction endonucleases occur as a complex with the methylase and a polypeptide that binds to the recognition site on DNA. They are often not very specific and cut at a remote site. Cuts nonspecifically at distance greater than 1000 bp from its recognition sequence.

Type II restriction endonucleases are the classic experimental tools. They have very specific recognition and cutting sites. The recognition sites are short, 4-8 nucleotides and are usually palindromic sequences. Because both strands have the same sequence running in opposite directions the enzymes make double stranded breaks, which, if the site of cleavage is off centre, generates fragments with short single

stranded tails, these can hybridise to the tails of other fragments and are called sticky ends.

They are generally named according to the bacterium from which they were isolated (first letter of genus name and the first two letters of the specific name). The bacterial strain is identified next and multiple enzymes are given roman numerals. For example the two enzymes isolated from the R strain of e. Coli are designated eco RI and eco RII. For eg. A class of endonucleases that cleaves DNA after recognizing a specific sequence, such as BamH1 (GGATCC), EcoRI (GAATTC), and HindIII (AAGCTT).

Type III. Cuts 24-26 bp downstream from a short, asymmetrical recognition sequence. Requires ATP and contains both restriction and methylation activities.

Making use of this logic/interesting facts in the data mining algorithm helps in performing step 3.

Step 5 : Identify Introns and Exons

The term "exon" is normally applied for regions which are not **spliced** out from a pre- mRNA sequence (**5' untranslated region (5' UTR), coding sequences (CDS)** and **3' untranslated region (3' UTR**). But this term is often used also to indicate the **protein- coding regions** only.

Exons contain the coding sequences of a gene - in contrast to introns, or "**junk DNA**," which are excised before **mRNA** is translated into a protein.

In eukaryotes, the gene is a combination of coding segments (**exons**) that are interrupted by non-coding segments (**introns**)

This makes computational gene prediction in eukaryotes even more difficult. Prokaryotes don't have introns - Genes in prokaryotes are continuous.

Regulatory regions: up to 50 kb upstream of +1 site Exons: protein coding and untranslated regions (UTR) 1 to 178 exons per gene (mean 8.8) 8 bp to 17 kb per exon (mean 145 bp)

Introns: splice acceptor and donor sites, junk DNA average 1 kb – 50 kb per intron

Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Splicing signals:

Exons are interspersed with introns and typically flanked by GT and AG

Intron1 Inton2

Exon 1 GT--------AG Exon2 GT-------AG Exon3

Adenine recognition site marks intron. .snRNPs binds around adenine recognition site The *spliceosome* thus forms and excises introns in the mRNA. The beginning and end of exons are signaled by donor and acceptor sites that usually have GT and AC dinucleotides. Detecting these sites is difficult, because GT and AC appear very often. Try to recognize location of splicing signals at exon-intron junctions. This has yielded a weakly conserved donor splice site and acceptor splice site Profiles for sites are still weak, and lends the problem to the Hidden Markov Model (HMM) approaches, which capture the statistical dependencies between sites. Witth these underlying facts, introns identification module can be incorporated in the data mining algorithm.

Step 6 : Identify Consensus Sequences

These are the sequences out of a series of DNA, RNA or proteins, that reflect the most common choice of base or amino acid at each position. Areas of particularly good agreement often represent conserved functional domains. The generation of consensus sequences has been subjected to intensive mathematical analysis.

Step 7: Identify Promoter/Terminal sequences/regions

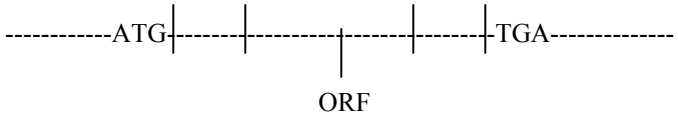
Promoter region : A region of DNA extending 150-300 bp upstream from the transcription start site that contains binding sites for RNA polymerase and a number of proteins that regulate the rate of transcription of the adjacent gene.

Terminator region : A DNA sequence that signals the end of transcription.

Identification of these regions is required in order to indicate start and end of the transcription process.

Step 8 : ORF identification

Open Reading Frame is a reading frame in a sequence of nucleotides in DNA that contains no termination codons and so can potentially translate as a polypeptide chain. Potential coding regions can be detected by looking at **ORFs.** Generally a genome of length *n* is comprised of (*n*/3) codons and Stop codons break genome into segments between consecutive Stop codons. The subsegments of these that start from the Start codon (ATG) are ORFs. ORFs in different frames may overlap.



Long open reading frames may be a gene. At random, we should expect one stop codon every (64/3) ~= 21 codons However, genes are usually much longer than this. A basic approach that can be used is to scan for ORFs whose length exceeds certain threshold This is naïve because some genes (e.g. some neural and immune system genes) are relatively short.

ORFs can be tested based on the codon usage as follows:

- Create a 64-element hash table and count the frequencies of codons in an ORF. Amino acids typically have more than one codon, but in nature certain codons are more in use.

- Uneven use of the codons may characterize a real gene
- This compensates for pitfalls of the ORF length test.

Testing ORF on the basis of Likelihood Ratio:

- An ORF is more "believable" than another if it has more "likely" codons.
- Do sliding window calculations to find ORFs that have the "likely" codon usage.
- Allows for higher precision in identifying true ORFs; much better than merely testing for length.
- **In-frame hexamer count** (frequencies of pairs of consecutive codons) method can also be used as further improvement.

Step 9: DNA to protein sequence conversion

As stated earlier, DNA to protein sequence conversion is not direct. It is DNA → transcription → RNA → translation → Protein process.

Ribonucleic Acid (RNA) is the messenger and is a temporary copy. It is never DNA → Protein because DNA is in nucleus and proteins are manufactured out of the nucleus. Hence a proofreading step is added. (Transcription = DNA→RNA). So actually… DNA → pre-mRNA → mRNA → Protein

Proteins carry out the cell's chemistry. These are large, complex molecules made up of smaller subunits called **amino acids.** They make up the cellular structure. There are 20 different **amino acids** whose different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.

RNA is first Translated to Protein, then Folded into the 3D structure. This is known as Protein Folding Problem.

3 base pairs of DNA/RNA(codon) → 1 amino acid. There can be 64 possible combinations of codons mapping to 20 amino acids. There is degeneracy of the genetic code i.e. several codons code for the same protein.

UAA, UAG and UGA correspond to 3 Stop codons that (together with Start codon ATG) delineate Open Reading Frames. Translation always starts with Methionine and ends with a stop codon. Approximately 10 codons are translated per second, but multiple translations can occur simultaneously. (Refer Fig. 5)

This information clearly indicates that the starting of protein synthesis process is with the start codon(ATG/AUG) and ends with any one of the stop codons(UAA/UAG/UGA) for one amino acid. Based on this logic, a module can be developed in the data mining algorithm using mRNA sequence as the input for the process.

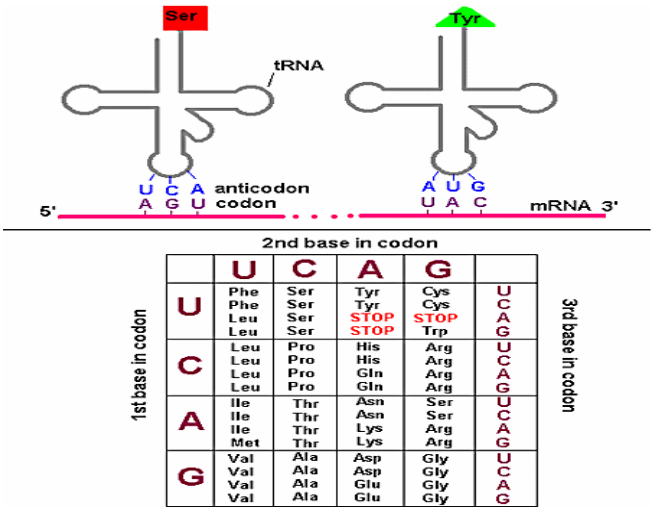


Fig. 5. The Genetic code and stop codons

V. ACKNOWLEDGEMENT

The authors gratefully acknowledge the contributions of S. Bajaj in the course of conceptualization of automation.

VI. REFERENCES

Papers from conference proceedings (Published) :
[1] D.Vassilev and J.Leunissen, "Application of Bioinformatics in Plant Breeding", in proc. 2005 Biotechnol & Biotechnol, pp.139-152.
Web Sites:
[2] www.bioalgorithms.info

VII. BIOGRAPHIES

Mrs. Manali Kshirsagar obtained her B.E. degree in Computer Technology from Nagpur University in the year 1992, and M.E. in Comp.Sc. & Engg. In 2001. She has presented papers in international and national conferences and also has publication in the international journal. She is an elected member of the Institution of Engineers(I), Nagpur Local Center. She is currently pursuing her doctoral programme in the area of Bioinformatics.

Prof (Col.) Gurmit Singh obtained his B. Tech. Degree in Electronics from College of Military Engineering; Pune (Maharastra); India & Military College of Telecommunication Engineering; Mhow (M.P.); India, in 1976; and M. Tech. Degree in Electrical Engineering from Indian Institute of Technology; Kanpur (U. P.); India in 1979.
Col. Singh has been Vice-Chairman of Computer Society of India, MHOW chapter and presented papers in the conventions of the society. Research work under his supervision has led to the award of three Doctor of Philosophy degrees. He has served in the Corps of Signals, Indian Army, from 1968 to 1993 and developed a Network Management System for a communication grid and also worked extensively in Real Time Decision Support Systems for Defence Forces.

Dr. G. Balasubramani, working as a Senior Scientist (Biotechnology) in Central Institute for Cotton Research (ICAR), Nagpur, Maharashtra. He did Bachelor degree in Agricultural Science from Tamil Nadu Agricultural University (TNAU), Coimbatore from 1984-88. He did his Master degree in Plant Biotechnology from the same University during 1988-90, the first batch of DBT sponsored program. He carried out research on *in vitro* nodulation and nitrogen fixation in *Sesbania rostrata*. He continued his Ph. D. program and carried out work on phylogenetic analysis based on DNA amplification & fingerprinting (DAF) in *Anabaena azollae*. He cleared the CSIR-UGC exam and awarded – Lectureship (NET) in 1991 + Fellowship for Ph. D. (Biotechnology) from 1991-1993. He also cleared ASRB / ICAR New Delhi NET (National Eligibility test) for Lectureship in 1992 and selected as Scientist in Plant Biotechnology. He has many national/international publications to his credit.