Unit 2

Functional Genomics:

* Study of how genes and intergenic regions of genome contribute to biological process
* Researcher studies genes on genome wide scale
* Goal is to determine how individual components work together
* Use current knowledge of gene function to develop a model
* Combines data derived from process related to DNA sequence, gene expression, etc.
* Several approaches:
  + DNA level
    - Genomics epigenomics
  + RNA level
    - Transcriptomics
  + Protein level
    - Proteomics
  + Metabolite level
    - Metabolomics

cDNA and cDNA generation:

* Complementary DNAs are double stranded DNA molecules
* One strand is DNA complementary to mRNA
* Other strand is mRNA sequence but T replaced by U
* Generation
  + Oligo(dt) primers are used to transform mRNA to cDNA
  + Oligo(dt) hybridize with poly A of mRNA
  + Reverse transcriptase is added, attaches to poly A tail and starts transcription
  + After the reverse transcription is done, product contains
    - 1mRNA
    - 1cDNA
  + RNAse is used to separate both
  + DNA polymerase I is used to exchange nucleotides with deoxynucleotide
  + Remaining strand is cut out
  + New fragment is synthesized by use of DNA polymerase I
  + Ligase enzyme is used to join DNA phosphodiester bonds
  + Double stranded cDNA is formed

What is clustering? Explain hierarchical clustering types

* Dividing data into number of groups
* Aim is to segregate groups with similar traits and assign them into clusters
* Hierarchical clustering
  + Algorithm that groups similar objects into groups called clusters.
* There are two types of hierarchical clustering
  + Divisive
    - 1 cluster containing the entire data set
    - Highest dissimilarity is reassignment its own cluster
    - Observations in old cluster similar to new cluster are assigned in another new cluster
    - Process repeats till all data in old cluster is in its own cluster
  + Agglomerative
    - Starts with each observation in own cluster
    - Two closest cluster are joined into one cluster
    - Close is defined by four metrics
      * Single linkage
        + Nearest neighbor
        + Shortest distance between pair
        + Sometimes produces clusters where observations from different clusters are closer than observations in the same cluster
      * Complete linkage
        + Farthest neighbor
        + Farther distance between pair
        + Produces tight clusters
        + Clusters can end up very close together
      * Average linkage
        + Distance between each pair is added and divided by number of pairs
        + Average and complete and two most popular linkage metrics
      * Centroid linkage
        + Distance between centroid of two clusters
        + Possible that smaller clusters are more similar than larges cluster
    - This continues until there is only one cluster containing whole dataset

ESTS

* Expressed Sequence Tags (ESTs) are short sequence reads
* Range of 100-700bp
* Obtained from randomly created cDNA clones
* Represents portions of expressed genes
* Present in database as cDNA/mRNA sequence
* Application of ESTs
  + Gene surveying
  + Gene identifications
  + Transcription mapping
  + Gene prediction
  + Quantification of gene expression
  + Reagents for downstream applications such as microarray and immuno screening
* Synthesis of ESTs
  + Isolation of mRNA from tissue
  + mRNAs are reverse transcribed to cDNA using oligo-DT primers
  + Generated cDNA separated by electrophoresis
  + Separated cDNAs will be selected on the basis of size and removed
  + Sequenced randomly from both ends
  + ESTs are generated and their sequences are compared using different databases

STS

* Sequence Tagged Sites STS is relatively short compared to ESTs
* PCR amplified sequence of 200 to 500 bp
* DNA sequence contains repetitive elements
* DNA primers complementary to the ends of the sequence is amplified using PCR
* Applications of STSs
  + Define unique, detectable landmarks
  + Server as markers for genetic and physical mapping of genes
  + Produces simple and reproducible pattern on agarose or polyacrylamide gel
* Synthesis of STSs:
  + Genome is broken up into fragments
  + Fragments replicated 10 times in bacterial
  + PCR is used to determine STS containing fragments
  + Primers bind to either side of STS so only the STS part is copied
  + Two fragments have same STS = overlapping parts of genome
  + Two different STSs = STSs must be near each other in the genome

SAGE:

* Serial Analysis of Gene Expression
* Global profiling of gene transcripts
* Requires prep of cDNA but does no require prior knowledge of genes
* Applications of SAGE:
  + Qualitative and quantitative assessment of every transcript present in a cell
  + Identification of differentially expressed transcripts
  + Comprehensive analysis of changes in mRNA
* Protocol for SAGE:
  + mRNA isolated and reverse transcribed
  + cDNA bound via biotin
  + cDNA cleaved
  + Cleaved DNA is washed out
  + Cleaved DNA tagged by removing beads via addition of oligonucleotides with sticky ends
  + Blunt ends tags are ligated to generate ditags
  + Ditags are cleaved to remove oligonucleotides
  + Transform concatemers into bacteria
  + Replication
  + Isolated concatemers form bacteria
  + Sequence