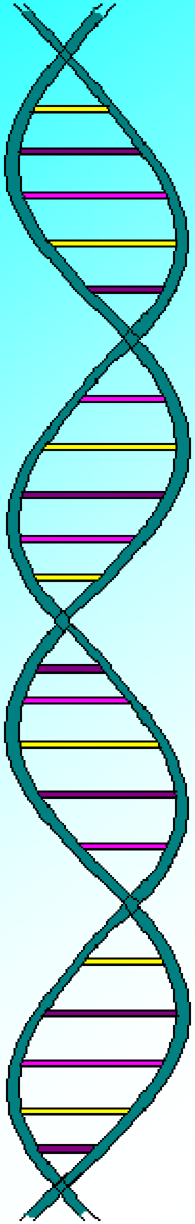
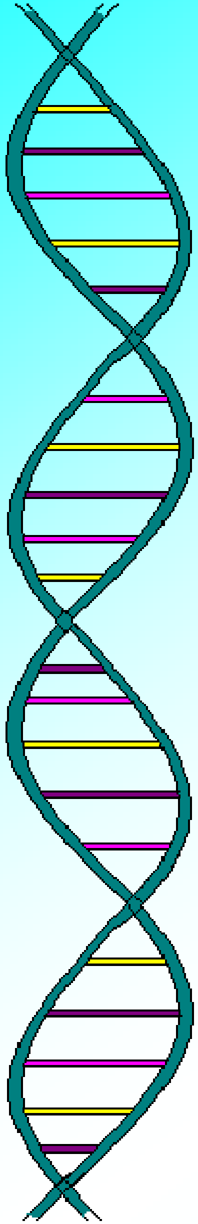


Presentation on

Overview of Bioinformatics

In this presentation……



Part 1 – Abbreviations

Part 2 – Foundations

Part 3 – Position of Bioinformatics

Part 4 – Methods in Bioinformatics

Part 5 – Extra slides

◁ ▷

Part 1

Abbreviations

- 2D-PAGE – two dimensional polyacrylamide gel electrophoresis
- BDGP – Berkeley Drosophila Genome Project
- CD – candidate drug
- cDNA – copy/complementary DNA
- CDS – coding sequence
- DDBJ – DNA Databank of Japan
- EBI – European Bioinformatics Institute
- EMBL – European Molecular Biology Laboratory

◀ ▶

- EP – expression profiler
- EST – expressed sequence tag
- ExPASy – Export Protein Analysis System (Switzerland)
- FN – false negative
- FP – false positive
- FRET – fluorescent resonance energy transfer
- GASP – Gene Annotation aSsessment Project
- GEO – gene expression omnibus
- GGTC – German Gene Trap Consortium

- GNOME – GNU network object model environment
- GOLD – Genomes Online Database
- GRAIL – gene recognition and assembly internet link
- HTG – high-throughput genomic sequence
- HTS – high-throughput screening
- KEGG – Kyoto encyclopedia of genes and genomes
- LCA – last common ancestor

◁ ▷

- LOG – Laplacian of Gaussian
- MAD – multi-wave length anomalous diffraction
- MAGE – microarray and gene expression
- MALDI – matrix-assisted laser desorption/ ionization
- ME – missing exon
- MGED – microarray gene expression database
- MIAME – minimum information about a microarray experiment

- MMDB – molecular modeling database
- mRNA – messenger RNA
- MS – mass spectrography/spectrometry
- MSD – macromolecular structure database
- NBRF – National Biomedical Research Foundation
- NCBI – National Centre for Biotechnology Information
- NDB – Nucleic acid Data Bank
- NMR – nuclear magnetic resonance

◀ ▶

- OMIM – Online Mendelian Inheritance in Man
- ORF – Open Reading Frame
- PAGE – polyacrlamide gel electrophoresis
- PAUP – phylogenetic analysis using parsimony
- PCR – polymerase chain reaction
- PDB – protein data bank
- PE – predicted exon
- PIR – protein information resource
- PN – predicted negative

◁ ▷

- PP – predicted positive
- RMSD – root mean square deviation
- rRNA – ribosomal RNA
- RT – reverse transcription
- SAGE – serial analysis of gene expression
- SELDI – surface-enhance laser desorption/ ionization
- SMART – simple modular architecture research tool
- SNP – single nucleotide polymorphism

- SMILES – simplified molecular input line entry specification
- SPR – surface plasmon resonance
- SRS – sequence retrieval system
- SSE – secondary structure element
- STS – sequence tagged site
- TE – true exon
- TN – true negative
- TP – true positive
- tRNA – transfer RNA
- WE – wrong exon

Computing related

- CGI – common gateway interface
- DBMS – database management system
- HMM – hidden Markov Model
- SQL – structured query language
- MAGE-ML – microarray gene expression markup language
- UML – unified modeling language
- XML – eXtensible Markup Language
- HTML – hypertext markup language
- PERL – practical extraction and reporting language

◁ ▷

- SOM – self-organizing map
- DNS – domain name server
- OOPS – object oriented programming system
- ISP – internet service provider
- TCP – transmission control protocol
- IP – internet protocol
- FTP – file transfer protocol
- HTTP – hypertext transfer protocol
- UNIX – unified information and computing service
- URL – uniform resource locator

Part

2

Foundations

Bioinformatics

The combination of biology and information technology. It is a branch of science that deals with the computer based analysis of large biological data sets. It incorporates the development of databases to store and search data, and of statistical tools and algorithms to analyze and determine relationships between biological sets, such as macromolecular sequences, structures, expression profiles and biochemical pathways

Components of Bioinformatics

- Creation of **databases** allowing storage and management of large biological data sets
- Development of **algorithms** and **statistics** to determine relationships among members of large data sets
- Use of these tools for the analysis and interpretation of various types of biological data, including DNA, RNA and protein sequences, protein structures, gene expression profiles and biochemical pathways

- Term “Bioinformatics” first came into use in the 1990s
- Computational tools for sequence analysis were available since 1960s
- Largely though not exclusively a computer-based discipline

Role of Bioinformatics

- Research on sequencing work being done since last two decades, but it is since two-three years that high density or large scale gene expression techniques for analysis of microarray data are being used
- Artificial Neural Network techniques are proving to be of greatest advantage in clustering/categorizing and analyzing them

Role of Bioinformatics **(contd…)**

- Experiments are producing large amounts of data
- Capable of providing insights into biological processes such as
  - Gene function
  - Gene development
  - Cancer
  - Aging
  - Pharmacology

- Bioinformatics will encompass and lead to other areas of study viz., comparative genomics, functional genomics, structural genomics and proteomics
- Other areas of informatics are computational biology, medical informatics, cheminformatics, genomics, proteomics, pharmacogenomics, etc.
- Bioinformatics involve areas such as creating repository databases, comparison of genes/ molecules/ proteins, expression and modeling, analysis from structures as well as sequences, predictions, etc.
- Vast and complex data generated by structural biologists using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy is done under bioinformatics

- Bioinformatics use certain disciplines of computer science such as AI, neural networks, genetic algorithms, dynamic programming, pattern recognition, etc.
- Quantitative (mathematical and/or computational) scientists talk about their interest in studying some aspect of “God’s mind”, whereas biologists are interested in “Mother Nature’s body”. If one needs to win Nature, one must be ready to meet her in the flesh
- Bioinformaticists can do popular work for biologists with proper collaboration with them

◁ ▷

- Bioinformatics is defined narrowly by some as the information science techniques needed to support genome analysis; many use it as synonymously with "computational molecular biology" or some even all of "computational biology"
- One of the most basic operations in bioinformatics involves searching for similarities, or homologies, between a newly sequenced piece of DNA and previously sequenced DNA segments from various organisms
- With all the variety of data, comes the potential for miscommunication. Getting various databases to talk to one another – what is called interoperability – is becoming more and more key as users flit among them to fulfill their needs. An obvious solution would be annotation – tagging data with names that are cross-referenced across databases and naming systems

- Bioinformatics is not a proper subset of biology or computer science, but is an interdisciplinary area of study combining both of them. There are two model of professionals coming in viz., computer scientists who specializes in biology and biologists who specializes in computer science
- Days have gone when scientists used to talk about computing capabilities of computers in general, especially the personal computers segment; it is time to think about high dimensional computing, grid computing, high performance clusters, powerful modeling software and many more
- Correspondingly, there is a need to think about changes in hardware and software configurations of the present day computers, as they would very soon become redundant. Watch for days in near future when a multi-processor computer would be available for the price of a personal computer…

- The aspirants have no further interest in pursuing courses in engineering, technology, management and other disciplines as most of them have already got saturated and reached the utmost. But the focus is shifting more towards biological sciences as there are many promising areas to work as well as numerous problems to solve
- In computing, there are many solutions available for data mining, pattern discover and machine learning, it is time for application of these techniques on the biological data and to develop tools and methods adaptable or suitable for this area
- There are no complete and integrated software tools are available at present for modeling DNA structure or prediction of molecules, eventhough it has been modeled many decades ago

- On one hand, the education, research and testing etc. aspects of drug design can be identified and brought under one roof so that it should be possible to provide everything right from beginning to end at one place. On the other hand, there is a necessity to consider the issues from different perspectives or dimensions such as academics, industry, government, and people
- The prime most of all difficulties eclipsing research and development in India in the bio-sciences discipline is lack of facilities such as microarray chips, protein chips, etc.

- The fruit fly (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), yeast (*Saccharomyces cerevisiae*) and the mouse are the four simple organisms used for sequencing genomes to study a variety of human diseases, including cancer and diabetes. It was observed that the proteins they encode closely resembles those of humans and are much easier to keep in laboratory
- Researchers found that 60 percent of the 289 known human disease genes have equivalents in flies and that bout 7,000 (50 percent) of all fly proteins show similarities to known mammalian proteins
- Researchers found that roughly one third (more than 6,000) of the worm's proteins are similar to those of mammals
- Researchers found that approximately 38 percent (about 2,300) of all yeast proteins could serve as a good model organism for studying cancer

- It was found that more than 90 percent of the mouse proteins identified so far show similarities to known human proteins
- Human and chimpanzee genomes are only 95 percent similar. Scientists compared 780,000 bp of chimps and human DNA found in GenBank
- In order to identify a DNA molecule, it would be sufficient to decode one strand of the two as both the strands are complementary to one another
- After the human genomes have been sequenced, the next step is to look at the messenger RNAs (mRNAs) and proteins. If DNA is the set of master blue prints a cell uses to construct proteins, then mRNA is like the copy of part of blueprint that a contractor takes to the building site every day. DNA remains in the nucleus of a cell; mRNAs transcribed from active genes leave the nucleus to give the orders for making proteins

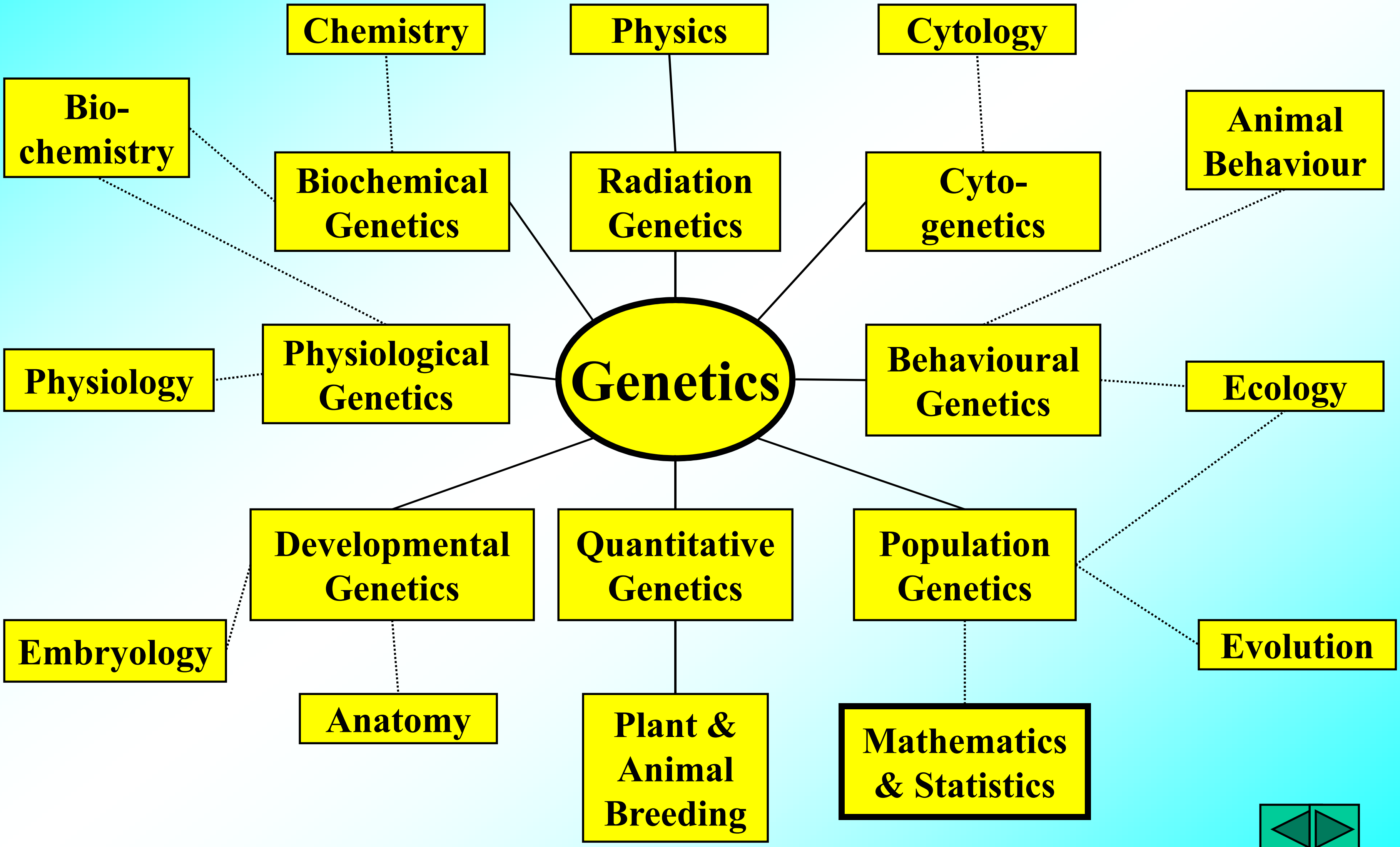
- Although every cell in the body contains all of the DNA code for making and maintaining a human being, many of those genes are never “turned on” or copied into mRNA, once embryonic development is complete. Various other genes are turned on and off at different times – or not at all – according to the tissue they are in and their role in the body. A pancreatic beta cell, for example, is generally full of the mRNA instructions for making insulin, whereas a nerve cell in the brain usually isn’t

- Earlier, scientists thought that one gene equals one mRNA equals one protein, but the reality is much more complicated. They now know that one gene can be read out in portions that are spliced and diced to generate a variety of mRNAs and that subsequent processing of the newly made proteins that those transcripts encode can alter their function. The DNA sequence of the human genome therefore tells only a small fraction of the story about what a specific cell is doing. Instead researchers must also pay attention to the transcriptome – the body of mRNAs being produced by a cell at any given time – and the proteome, all the proteins being made according to the instructions in those mRNAs

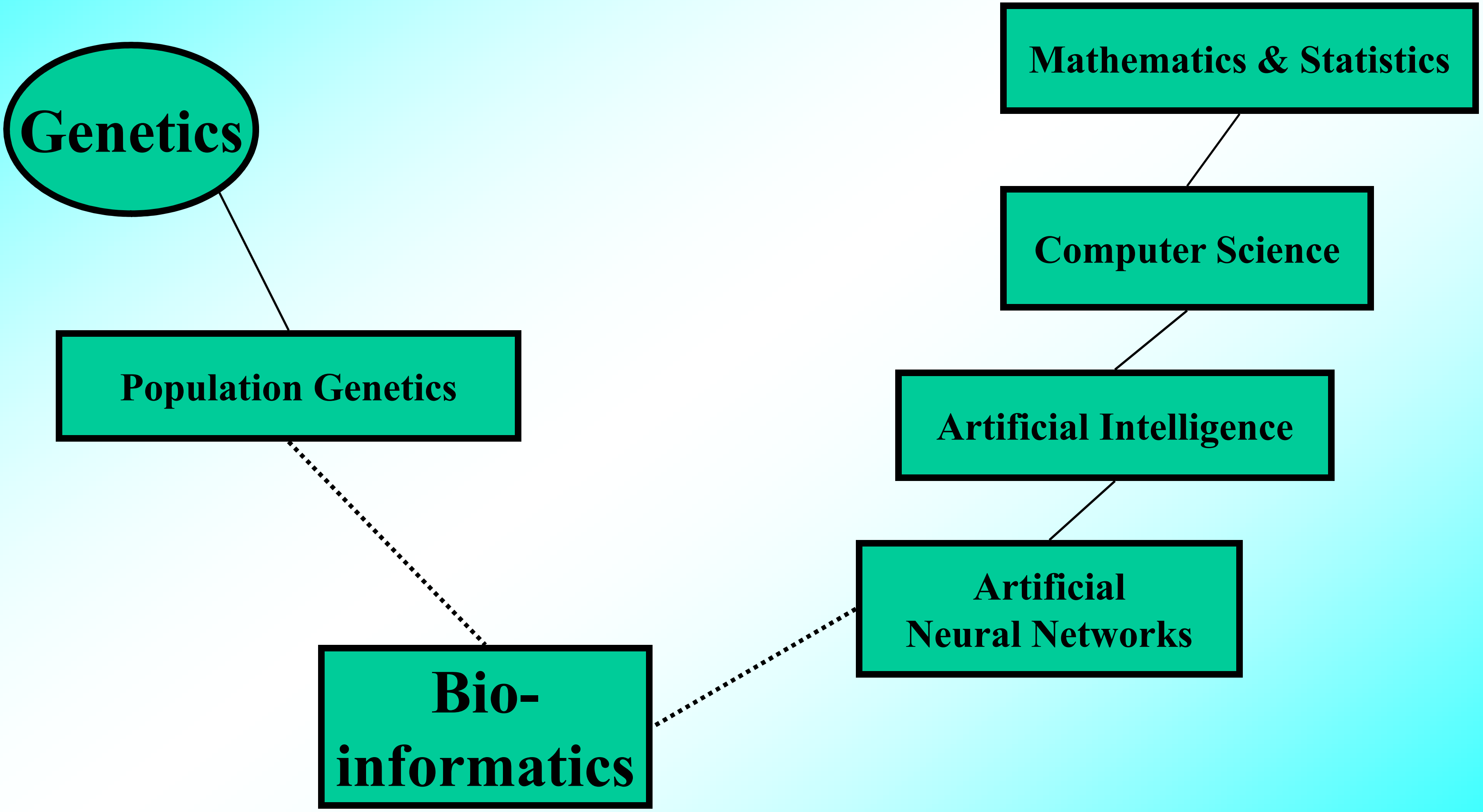
Part 3

Position of Bioinformatics

Genetics and related fields



Position of Bio-informatics



Genetics

Population Genetics

Bio-informatics

Mathematics & Statistics

Computer Science

Artificial Intelligence

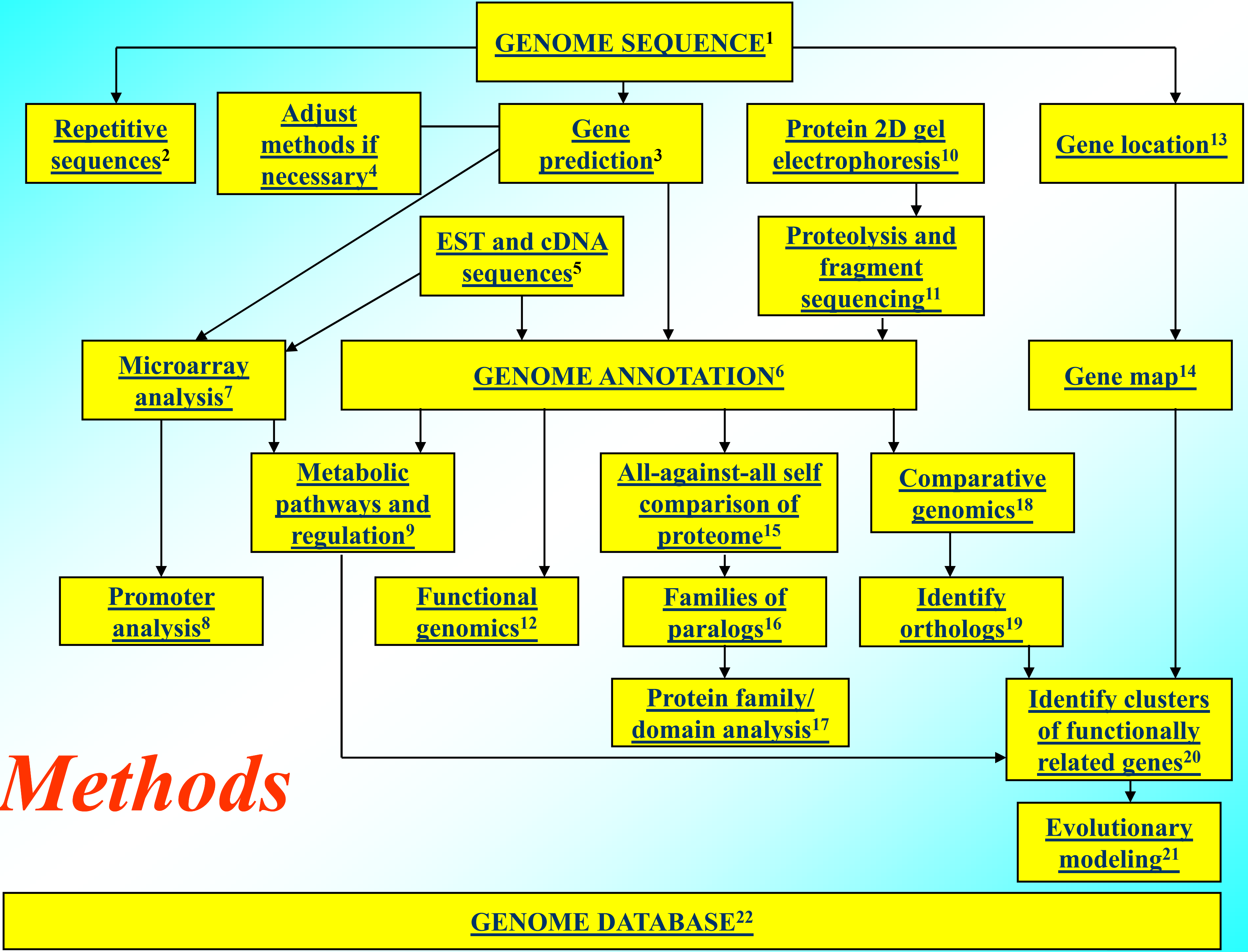
Artificial Neural Networks

Part

4

Methods in Bioinformatics

Methods



GENOME SEQUENCE[1]

Repetitive sequences[2]

Adjust methods if necessary[4]

Gene prediction[3]

Protein 2D gel electrophoresis[10]

Gene location[13]

EST and cDNA sequences[5]

Proteolysis and fragment sequencing[11]

Microarray analysis[7]

GENOME ANNOTATION[6]

Gene map[14]

Metabolic pathways and regulation[9]

All-against-all self comparison of proteome[15]

Comparative genomics[18]

Promoter analysis[8]

Functional genomics[12]

Families of paralogs[16]

Identify orthologs[19]

Protein family/ domain analysis[17]

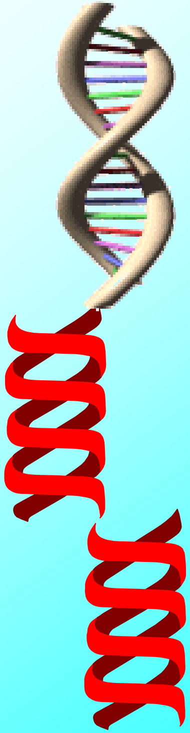
Identify clusters of functionally related genes[20]

Evolutionary modeling[21]

GENOME DATABASE[22]

End of Presentation...

Thank you....



Genome Sequence [1]

Genome sequences are assembled from DNA sequence fragments of approximate length 500 bp obtained using DNA sequencing machines. Chromosomes of a target organism are purified, fragmented, and sub-cloned in fragments of size hundreds of kbp in bacterial artificial chromosomes (BACs). They are then further sub-cloned as smaller fragments into plasmid vectors for DNA sequencing. Full chromosomal sequences are then assembled from the overlaps in a highly redundant set of fragments by an automatic computational method (Myers et al. 2000) or from a fragment order on a physical map.

Repetitive Sequences [2]

Eukaryotic genomes comprise classes of repeated elements, including tandem repeats present in centromeres and telomeres, dispersed tandem repeats (minisatellites and macrosatellites), and interdispersed TEs. TEs can comprise one-half or more of the genome sequence. Analysis of sequence repeats and identification of classes of repeated elements is aided by searchable databases

Gene Prediction [3]

Gene identification in prokaryotic organisms is simplified by their lacking introns. Once the sequence patterns that are characteristics of the genes in a particular prokaryotic organism (e.g. codon usage, codon neighbour preference) have been found, gene locations in the genome sequence can be predicted quite accurately. The presence of introns in eukaryotic genomes makes gene prediction more involved because, in addition to the above features, locations of intron-exon and exon-intron splice junctions must also be predicted. Methods for gene prediction in prokaryotes and eukaryotes are available.

Adjust methods if necessary [4]

Gene identification methods involve training a gene model (e.g. a hidden Markov model or neural network) to recognize genes in a particular organism. Due to variations in gene codon preferences and splice junctions, a model must usually be trained for each new genome.

⮍ ◁ ▷

EST and cDNA Sequences [5]

Since gene prediction methods are only partially accurate (Bork 1999), gene identification is facilitated by high-throughput sequencing of partial cDNA copies of expressed genes (called expressed sequence tags or EST sequences). Presence of ESTs confirms that a predicted gene is transcribed. A more thorough sequencing of full-length cDNA clones may be necessary to confirm the structure of genes chosen for a more detailed analysis.

Genome Annotation [6]

The amino acid sequence of proteins encoded by the predicted genes is used as a query of the protein sequence databases in a database similarity search. A match of a predicted protein sequence to one or more database sequences not only serves to identify the gene function, but also validates the gene prediction. Pseudogenes, gene copies that have lost function, may also be found in this analysis. Only matches with highly significant alignment scores and alignments should be included. The genome sequence is annotated with the information on gene content and predicted structure, gene location, and functional predictions. The predicted set of proteins for genome is referred to as the proteome. Accurate annotation is extremely important so that other users of information are not misinformed. Usually, not all query proteins will match a database sequence. Hence, it is important to extend the analysis by searching the predicted protein sequence for characteristic domains (conserved amino acid patterns that can be aligned) that serve as a signature of a protein family or of a biochemical or structural feature. A further extension is to identify members of protein families or domains that represent a structural fold using various computational tools. This additional information also needs to be accurately described and significance established.

Microarray Analysis [7]

Microarray analysis provides a global picture of gene expression for the genome by revealing which genes are expressed at a particular stage of the cell cycle or developmental cycle of an organism, or genes that respond to a given environmental signal to the same extent. This type of information provides an indication as to which genes share a related biological function or may act in the same biochemical pathway and may thereby give clues that will assist in gene identification.

⮌ ◁ ▷

Promoter Analysis [8]

Genes that are found to be coregulated either by a microarray analysis or by a protein 2D analysis should share sequence patterns in the promoter region that direct the activity of transcription factors. There are many types of analyses performed, and a number of tools available for analyzing coregulated genes

Metabolic Pathways and Regulation [9]

As genes are identified in a new genome sequence, some will be found that are known to act sequentially in a metabolic pathway or to have a known role in gene regulation in other organisms. From this information, the metabolic pathways and metabolic activities of the organism will become apparent. In some cases, the apparent absence of a gene in a well-represented pathway may lead to a more detailed search for the gene. Clustering of genes in the pathway on the genome of a related organism can provide a further hint as to where the gene may be located

Protein 2D Gel Electrophoresis [10]

Individual proteins produced by the genome can be separated to a large extent by this method and specific ones identified by various biochemical and immunological tests. Moreover, changes in the levels of proteins in response to an environment signal can be monitored in much the same way as a microarray analysis is performed. Microarrays only detect untranslated mRNAs, whereas 2D gel protein analysis detects translated products, thus revealing an additional level of regulation.

Proteolysis and Fragment Sequencing [11]

Protein spots may be excised from a 2D protein gel and subjected to a combination of amino acid sequencing and cleavage analyses using the techniques of mass spectrometry and high-pressure liquid chromatography. Genome regions that encode these sequences can then be identified and the corresponding gene located. A similar method may be used to identify the gene that encodes a particular protein that has been purified and characterized in the laboratory.

⮍ ◁ ▷

Functional Genomics [12]

Functional genomics involves the preparation of mutant or transgenic organisms with a mutant form of a particular gene usually designed to prevent expression of the gene. The gene function is revealed by any abnormal properties of the mutant organism. This methodology provides a way to test a gene function that is predicted by sequence similarity to be the same as that of a gene of known function in another organism. If the other organism is very different biologically (comparing a predicted plant or animal gene to a known yeast gene), then functional genomics can also shed light on any newly acquired biological role. When two or more members of a gene family are found, rather than a single match to a known gene, the biological activity of these members may be analyzed by functional genomics to look for diversification of function in the family.

Gene Location [13]

Since the entire genome sequence is available, as each gene is identified, the relative position of the gene will be known.

Gene Map [14]

A map showing the location of each identified gene is made. These positions of genes can be compared to similar maps of other organisms to identify rearrangements that have occurred in the genome. Gene order in two related organisms reflects the order that was present in a common ancestor genome. Chromosomal breaks followed by a reassembly of fragments in a different order can produce new gene maps. These types of evolutionary changes in genomes have been modeled by the chromosome, but also by genetic analysis. Populations of an organism show sequence variations that are readily detected by DNA sequencing and other analysis methods. The inheritance of genetic diseases in humans and animals (e.g., cancer and heart disease), and of desirable traits in plants, can be traced genetically by pedigree analysis or genetic crosses. Sequence variations (polymorphisms) that are close to (tightly linked) a trait may be used to trace the trait by virtue of the fact that the polymorphism and the trait are seldom separated from one generation to the next. These linked polymorphisms may then be used for mapping and identifying important genes

↺ ◀ ▶

All-against-all Self Comparison of Proteome [15]

A comparison is made in which every protein is used as a query in a similarity search against a database composed of the rest of the proteome, and the significant matches are identified by a low expect value ($E<10^{-6}$ was used in a recent analysis by Rubin et al., 2000). Since many proteins comprise different combinations of a common set of domains, proteins that align along most of their lengths (80 %) identity is a conservative choice) re chosen to select those that have a conserved domain structure.

Families of Paralogs [16]

A set of related proteins identified in step 15 is subjected to a cluster analysis in order to identify the most closely related groups of proteins and to avoid domain-matching. This group of proteins is derived from a gene family of paralogs that have arisen by gene duplication.

Protein Family/Domain Analysis [17]

Each protein in the predicted proteome is again used as a query of a curated protein sequence database such as SwissProt in order to locate similar domains and sequences. The domain composition of each protein is also determined by searching for matches in domain databases such as Interpro. The analysis reveals how many domains and domain combinations are present in the proteome, and reveals any unusual representation that might have biological significance. The number of expressed genes in each family can also be compared to the number in other organisms to determine whether or not there has been an expansion of the family in the genome.

Comparative Genomics [18]

Comparative genomics is a comparison of all the proteins in two or more proteomes, the relative locations of related genes in separate genomes, and any local grouping of genes that may be of functional or regulatory significance.

⮍ ◀ ▶

Identify Orthologs [19]

Orthologs are genes that are so highly conserved by sequence in different genomes that the proteins they encode are strongly predicted to have the same structure and function and to have arisen from a common ancestor through speciation. To identify orthologs, each protein in the proteome of an organism is used as a query in a similarity search of a database comprising the proteomes of one or more different organisms. The best hit in each proteome is likely to be with an ortholog of the query gene. In comparing two proteomes, a common standard is to require that for each pair of orthologs, the first of the pair is the best hit when the second is used to query the proteome of the first. To find orthologs, very low E value scores ($E < 10^{-20}$) for the alignment that includes 60-80% of the query sequence are generally required in order to avoid matches to paralogs. Although these requirements for classification of orthologs are very stringent, a more relaxed set of conditions will lead to many more false-positive predictions. In bacteria, the possibility of horizontal transfer of genes between species also has to be considered.

↺ ◀ ▶

Identify Clusters of Functionally Related Genes [20]

In related organisms, both gene content of the genome and gene order on the chromosome are likely to be conserved. As the relationship between the organisms decreases, local groups of genes remain clustered together, but chromosomal rearrangements move the clusters to other locations. In microbial genomes, genes specifying a metabolic pathway may be contiguous on the genome where they are coregulated transcriptionally in an operon by a common promoter. In other organisms, genes that have a related function can also be clustered. Hence, the function of a particular gene can sometimes be predicted, given the known function of a neighbouring, closely linked gene. Genomes are also compared at the level of gene content, predicted metabolic functions, regulation as revealed by microarray analysis, and others. These comparisons provide a basis for additional predictions as to which genes are functionally related. Gene fusion events that combine domains found in two proteins in one organism into a composite protein with both domains in a second organism are also found and provide evidence that the protein physically interact or have a related function.

Evolutionary Modeling [21]

Evolutionary modeling can include a number of types of analyses including

(1) the prediction of chromosomal rearrangements that preceded the present arrangement (e.g., a comparison of mouse and human chromosome)

(2) analysis of duplications at the protein domain, gene, chromosomal, and full genome level

(3) search for horizontal transfer events between separate organisms

Genome Database [22]

Due to the magnitude of the task, the earlier stages of genome analysis including gene prediction and database similarity searches are performed automatically with little human intervention. The genome sequence is then annotated with any information found without involving human judgment. The types of genome analyses in the flowchart also provide many predictions and give rise to many preliminary hypotheses regarding gene function and regulation. As more detailed information is collected by laboratory experiment and by a closer examination of the sequence data, this information needs to be linked to the genome sequence. In addition, the literature, past and present needs to be scanned for information relevant to the genome. A carefully crafted database that takes into account the entire body of information should then be established. In addition to information on the specific genome of interest, the database should include cross-references to other genomes. To facilitate such intergenome comparisons, common gene vocabularies have been proposed. This slow, expensive, and time-consuming phase of genome analysis is of prime importance if the genome information is to be available in an accurate form for public use.

Part 5

Extra Slides

- SNPs
- Databases
- Protein structure
- Protein engineering
- Definitions
- Clinical implications
- Genetic algorithms

◁ ▷

- Sequence analysis
- Sequence alignment
- Protein coding regions
- Phylogeny and Phylogenetic trees
- Dayhoff and BLOSUM matrix
- Protein folding and fold recognition
- Protein stability, hydrophobicity, structures
- 2D gel electrophoresis
- Horizontal gene transfer

- Context free grammars for modeling RNA secondary structure
- Multiple sequence analysis
- Phylogenetic comparison and prediction
- Alpha helix, Beta sheet, Loop, Coil
- Microarray technology
- Mass spectrography
- Dendograms
- Hidden Markov model
- Bayesian clustering

- Evolutionary Computing (EC) = Genetic Algorithms (GA) + Evolution Strategies (ES) + Evolutionary Programming (EP)
- Soft Computing (SC) = Evolutionary Computing (EC) + Artificial Neural Networks (ANN) + Fuzzy Logic (FL)

◁ ▷