Code: AS-2908

Integrated UG/PG Biotechnology (Fifth Semester)
End Semester Examination, 2013
LBTC 504: Bioinformatics

Model Answers

Answer Key for Section-A (Objective Type Questions)

Question 1.

(i) (c) Both of the above
(ii) (a) Performed on computer
(iii) (c) Both a and b
(iv) (c) A symbol can be aligned with a gap
(v) (d) All of the above
(vi) (c) BLAST
(vii) (b) Genome
(viii) (d) Both a and b
(ix) (d) All of the above
(x) (a) First pH, then Molecular Weight

Model Answers for Section-B (Long Answer Type Questions)

Answer 2:

(i) Bioinfomatics is a recent scientific discipline for the computational analysis and storage of biological data. The word bioinformatics is derived from two words: Bio means biology and Informatique meaning ‘data processing’. The domain of biology, computer science and information technology is used for managing and analyzing biological data using advanced computing techniques. The knowledge of computer science and information technology is applied for creation as well as management of databases, data warehousing, data mining and communication networking. Bioinformatics can be defined as “Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyze or visualize such data”. Bioinformatics has emerged as a very promising and important discipline for academics, research and industrial application for the data storage, data warehousing and analyzing the sequences that has influenced scientific, engineering and economic development of the world.

Bioinformatics is called a multidisciplinary subject as in bioinformatics, knowledge of many branches is required. Basically, it combines biology, computer science, information technology, physics, mathematics, statistics, etc. and has powerful impact on all disciplines of life science, and hence is characterized by its multidisciplinary approach. In parallel to bioinformatics, several related and partially overlapping research areas have evolved. Computational biology focuses more on computations and simulations associated with biological macromolecules and biological systems, including simulation of cellular processes. Mathematical biology analyses e.g. population dynamics in complex ecological systems, whereas physical biology studies physical processes and phenomena found in biological systems. Nanobiotechnology combines nanomaterials and nanotechniques with biotechnology and biological molecules, whereas nanomedicine uses nanotechnology for medical purposes. Medical informatics handles and retrieves medical data, e.g. electronic medical records or medical publications. Medical statistics and statistical genetics uses statistical methods to analyse medical data looking for correlation between genetic variation and susceptibility to specific diseases.

(ii) Major applications of Bioinformatics are as follows:

1. Sequence Analysis

- **(a) Nucleotide sequence analysis:** DNA of various organisms is sequenced and stored as databases for easy retrieval and comparison. Following sequence analysis, alignment of two sequences is done to determine the parts of the sequences conserved from one species to the next and study the divergence of an organism from other organisms in evolution.
- **(b) Sequence translation:** If we know the nucleotide sequence, then it can be converted into amino acid sequence or vice versa.
- **(c) Protein sequence analysis:** After we get the amino acid sequence of a protein, several tools could be used to do further analysis of the protein like molecular weight, isoelectric point, titration curves, hydrophobicity, etc.
- **(d) Structure analysis:** Specific programs could be used to visualize the 3-D shape of proteins and nucleotides that gives greater understanding of the structure and function.

2. In silico prediction of drugs or Pharmacoinformatics

- **(a) Pharmacoinformatics:** It refers to implementation and use of computational tools and information technologies for the discovery and development of drugs. It involves two major components viz. Pharmacogenomics (genomic level) and Pharmacogenetics (genetic level). Pharmacogenomics is the technology that analyses how genetic makeup of an individual affects response to drugs. It deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms (SNPs) with a drug's efficacy or toxicity. By doing so, pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects. Such approaches will lead to the development of “personalized medicine"; in which drugs and drug combinations are optimized for each individual's unique genetic makeup. This will lead to decreased incidences of adverse drug reactions and no-response to various drug formulations.
- **(b) Gene therapy:** It is the use of DNA as a pharmaceutical agent to treat disease by using DNA that encodes a functional, therapeutic gene to replace a mutated gene, directly correcting a mutation and using DNA that encodes a therapeutic protein drug.
- **(c) Antibiotic resistance:** Prolonged use of antibiotics leads to resistance against antibiotics in bacterial population. By using bioinformatics tools, genetic markers could be developed to detect pathogenic strains which could help to control the spread of infections.

3. **Agroinformatics:** Agroinformatics refers to the application of bioinformatics in the agricultural science including plant genomes. By knowing the genomes of plants, we can generate drought resistance varieties, soil alkalinity resistant varieties, abiotic stress tolerating varieties. Insect resistant varieties can also be developed using comparative plant genomics. Example, Bt cotton. This will lead to a decrease in the amount of insecticides usage which will lead to an increase in nutritional quality and a decreased cost of production resulting in an increased farm productivity and help in getting food security. Similarly, sequencing projects of farm animals like cow, pigs, sheep, etc. can help in enhancing the understanding of biology of these animals which will help in increasing the production of milk, meat, etc. This will lead to increased nutrition and better health of our livestocks.

4. **Phylogenomics and Evolutionary Studies:** Phylogenomics is the use of phylogenetic methods to predict protein function via evolutionary analysis of a gene and its homologs. Information on structural, functional and comparative analysis of these genomes and genes from a wide variety of organisms will be useful to understand more about their evolution. PHYLIP is one of the most popular bioinformatics tool for conducting phylogenetic analysis.

5. **Literature Retrieval:** Bioinformatics tools help scientists and researchers to search various databases for information which helps in designing their projects and identifying the research that has been done or in progress. The literature/citation database at NCBI called PubMed can be easily searched with Entrez by a simple keyword search.

(iii) A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity between the sequences by subjecting them to various sequencing methods. When two symbolic representations of DNA or protein sequences are arranged next to one another so that their most similar elements are juxtaposed they are said to be aligned. In a symbolic sequence, each base or residue monomer in each sequence is represented by a letter. The convention is to print the single-letter codes for the constituent monomers in order in a fixed font (from the N-most to C-most end of the protein sequence in question or from 5' to 3' of a nucleic acid molecule). This is based on the assumption that the combined monomers evenly spaced along the single dimension of the molecule's primary structure. Every element in a trace is either a match or a gap. Where a residue in one of two aligned sequences is identical to its counterpart in the other the corresponding amino-acid letter codes in the two sequences are vertically aligned in the trace: a match. When a residue in one sequence seems to have been deleted since the assumed divergence of the sequence from its counterpart, its "absence" is labelled by a dash in the derived sequence. When a residue appears to have been inserted to produce a longer sequence a dash appears opposite in the unaugmented sequence. Since these dashes represent "gaps" in one or other sequence, the action of inserting such spacers is known as gapping. A deletion in one sequence is symmetric with an insertion in the other. When one sequence is gapped relative to another a deletion in sequence a can be seen as an insertion in sequence b. The two types of mutation are referred to together as indels.

Sequence alignment helps in:

(a) The comparison of sequences to find homology (common ancestor sequence).
(b) Identifying the same or related structure and function.
(c) Identifying the relationships between the differences in alignments and functional changes or disease.

An alignment of two sequences must satisfy the following conditions:

(a) Two gaps or blanks, written as '-' cannot be aligned.
(b) We can align one symbol (character) from sequence A with one symbol from sequence B (either a match or a mismatch).
(c) A symbol present in sequence A or in sequence B can be aligned with a gap.
(d) All symbols present in sequence A and sequence B have to be in the alignment and in the same order as they appear in the unaligned sequences.

(iv) Substitution matrices are the scoring matrices based on the observed substitution rates among various sequences present in nature. A substitution matrix describes the rate at which one character in a sequence changes to other character over time. Substitution matrices are usually seen in the context of amino acid or DNA sequence alignments, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix.

Substitution matrices are of two types:

(a) PAM (Point Accepted Mutation) matrix was developed by Margaret Dayhoff in the 1978. In these matrices, substitutions of amino acids are observed in homologous protein sequences during evolution, so that these amino acid substitutions do not significantly change the function of the protein. Hence, these mutations are accepted by natural selection. To prepare PAM matrices, observed mutations in alignments between similar sequences were estimated and then used to generate a 20X20 mutation probability matrix P representing all possible amino acid changes. PAM1 means 1 point mutation per 100 amino acids. By multiplying PAM1 by itself, we can derive matrix with multiple PAM units to approximate the substitution rates over multiple PAM units. PAM250 is widely used because it provides a better alignment score for distantly related proteins of about 20% amino acid identity.
(b) The BLOSUM (BLOck SUbstitution Matrix) was developed by Steven Henikoff and Jorja G. Henikoff in 1992 from conserved regions called blocks, derived from the Blocks database. The Blocks database contains multiple aligned, ungapped segments corresponding to the highly conserved regions of proteins. Steps to prepare BLOSUM Matrix:

1. Preparation of frequency table of amino acids from database of blocks.
2. The table is used to calculate a matrix representing the odds ratio between these observed frequencies and those expected by chance.

BLOSUM 62 is derived from Blocks containing >62% identity in ungapped sequence alignment. It is the most suitable matrix for aligning protein sequences in comparison to other BLOSUM as well as PAM Marices. BLOSUM 62 is the default matrix for the standard protein BLAST program.

(v) **Biological Data Analysis:** It can be defined as the application of statistical and other analytical tools for interpreting and correlating the various components of data.

Step-by-step analysis of biological data:

1) Specification of the biological question.
2) Put the question in the form of a biological null hypothesis and alternate hypothesis.
3) Put the question in the form of a statistical null hypothesis and alternate hypothesis.
4) Determine which variables are relevant to the question.
5) Determine what kind of variable each one is (An independent variable, sometimes called an experimental variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable, sometimes called an outcome variable, Categorical variables are also known as qualitative variables, Continuous variables are also known as quantitative variables. Continuous variables can be further categorized as either interval or ratio variables).
6) Design an experiment that controls or randomizes the confounding variables.
7) Based on the number of variables, the kind of variables and the hypothesis to be tested, choose the best statistical test to use.
8) If possible, do a power analysis to determine a good sample size for the experiment.
9) Do the experiment.
10) Apply the statistical test you chose, and interpret the results.
11) Communicate your results effectively, usually with a graph or table.

Sequence of Biological Data Analysis

Data collection-Classification-Assignment of test site-Reference condition for comparison-Assessment

Applications of Biological data analysis:

(a) Comparing microbial typing method
(b) Comparing regulatory network distance function
(c) Comparing the agreement of different gene expression data set with gene pathway information
(d) Microarray Data Analysis
(e) Real-time PCR Data Analysis (delta delta Ct method)
(f) Protein Expression Data Analysis (Densitometry)

(vi) Gene Prediction is used for identifying the coding region for a protein, finding the open reading frame, the start and end of a gene, finding the exon-intron boundaries in eukaryotes and finding the regulatory sequences for a gene. For prokaryotes, gene prediction is relatively easy compared to eukaryotes as they have small genomes, high coding density (>90%) and no introns. Therefore, gene identification becomes relatively easy with success rate ~ 99%. However, the problems are overlapping ORFs, short genes and finding TSS and promoters. In eukaryotes, the large genome size, low coding density (<50%) and presence of intron/exon structure make gene identification a complex problem with a gene level accuracy of about 50%.

Approaches for gene prediction

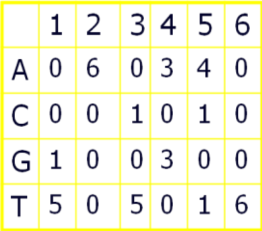
1. **Similarity-based methods (extrinsic)**: use similarity to annotated sequences for proteins, cDNAs, ESTs, based on sequence conservation due to functional constraints, use local alignment tools (Smith-Waterman algorithm, BLAST, FASTA) to search protein, cDNA, and EST databases, will not identify genes that code for proteins not already in databases (can identify ~50% new genes), limits of the regions of similarity not well defined.
2. **Comparative genomics:** Based on aligning of genomic sequences from different species and on the assumption that coding sequences are more conserved than non-coding. Two approaches are used for comparative genomics:
   Intra-genomic (gene families) and Inter-genomic (cross-species)
   Alignment of homologous regions is done but it is difficult to define limits of higher similarity.
3. ***Ab initio* gene-finding (intrinsic):** This method is used to identify only coding exons of protein coding genes by integrating coding statistics with signal detection (signal sensors).
   Coding statistics: It is a function that computes the likelihood for a given DNA sequence whether the sequence is coding for a protein or not. We can use unequal usage of codons in the coding regions to differentiate between coding and non-coding regions of the genome.
   Signal Sensors: Various pattern recognition methods are used for identification of signals which are a string of DNA recognized by the cellular machinery. Some examples of signal sensors are:

(a) Consensus sequences: obtained by choosing the most frequent base at each position of the multiple alignment of subsequences of interest

```
                  TACGAT
                  TATAAT
                  TATAAT
                  GATACT
                  TATGAT
                  TATGTT
Consensus sequence TATAAT
```

(b) Positional weight matrix: computed by measuring the frequency of every element of every position of the site (weight).

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT



	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

Score for any putative site is the sum of the matrix values (converted in probabilities) for that sequence (log-likelihood score).

Disadvantages: Cut-off value required and it assumes independence between adjacent bases.

(c) **Markov Model:** A process is Markov if it has no memory, that is, if the next state it assumes, depends only on its present state and not on any previous states. The states can be observed and the transition probabilities between states can be computed.

4. **Integrated approaches:** These approaches integrate results from different sources. Example of programs that integrate results of similarity searches with *ab initio* techniques (GenomeScan, FGENESH+, Procrustes), programs that use synteny between organisms (ROSETTA, SLAM) and integration of programs predicting different elements of a gene (EuGène).

(vii) **Microarrays:** A high throughput technology that allows detection of thousands of genes simultaneously to provide a global view on biological processes. The probe sequences are designed and placed on an array in a regular pattern of spots. The chip or slide is usually made of glass or nylon and is manufactured using technologies developed for silicon computer chips. Each microarray chip is arranged as a checkerboard of 10^5 or 10^6 spots or features, each spot containing millions of copies of a unique DNA probe (often 25 nt long). Microarrays use hybridization to detect a specific DNA or RNA in a sample by using millions of different probes, fixed on a solid surface, to probe complex DNA mixture. The exact sequence of the probes at each feature/location on the chip is known. Wherever some of the sample DNA hybridizes to the probe in a particular spot, the hybridization can be detected because the target DNA is labeled (and unbound target is washed away). Therefore one can determine which of the million different probe sequences are present in the target. The amount of signal directly depends on the quantity of labeled target DNA. Thus microarrays can give a quantitative description of how much of a particular sequence is present in the target DNA. This is particularly useful for studying gene expression, one common application of microarray technology.

Features of microarrays:

(a) Parallelism: Analysis of thousands of genes simultaneously
(b) Miniaturization: Small chip size
(c) Multiplexing: Multiple samples at the same time
(d) Automation: Chip manufacturing, Reagents

Applications:

(a) Study of expression of genes over time, between tissues, and disease states
(b) Identification of complex genetic diseases
(c) Drug discovery and toxicology studies
(d) Mutation/polymorphism detection (SNPs)
(e) Pathogen analysis

Overview of a microarray experiment

1. **Array fabrication:** This step includes the microarray fabrication using either the presynthesized probes or in situ synthesis of probes.
2. **Target/Sample preparation and labelling:** The target is labelled with either radioactivity (P^{33}, S^{53}, H^{3}) or tagged with fluorescence dyes (Cy5 red dye, Cy3 green dye) for sensitive detection.
3. **Hybridization of labelled targets with immobilized probes:** The sequence complementarity forms the basis of hybridization between labelled target and immobilized probe and microarray technology relies on hybridization for expression analysis.
4. **Detection of expression level:** Detection of hybridized targets on probe depends on the label used for the target. For radioactively labelled targets, phosphorimaging systems are used. For fluorescently labelled targets, the arrays are irradiated and fluorescence intensities are detected. Confocal microscopy or a CCD camera is used for taking fluorescent images.
5. **Data analysis:** The data generated by microarray experiments are hybridized images. These images are analyzed to identify differentially and co-expressed genes. The gene expression data from microarray experiments are generated in the form of a matrix. In this data matrix, each row represents a gene and each column represents the sample.