Your NAME: Asma Ishaq Tahir Your Matrikel-Number: 2853680

Today's date: 09-Feb-2016

Max number of points: 100

> 95	→ 1,0
86 – 95	→ 2,0
76 – 85	→ 3,0
66 – 75	→ 4,0
< 66	→ 5,0

(1) What are the three data formats or data tiers of ENA (European Nucleotide Archive)? *(3 points)*

The three data formats of ENA are:

→ Text
→ XML
→ FASTA

These type of files can be downloaded from ENA entry.

(2) Provide 4 examples for genome databases and describe briefly the content and structure of these databases. *(4 points)*

Example of 4 genome databases are:

① ENSEMBL: Ensembl is the most widely used genome database. It have the genomes of human, mouse + zebra fish. Highly annotated, reviewed, non-redundant database. Provides information about gene, protein, transcripts, variations, comparative genomics, DNA-Seq data + regulatory regions.

② MOUSE GENOME INFORMATICS: It is the database of mouse genome. It connects the information from the human genome also to show the similarities between their genes, proteins. It display the information of gene, protein, sequence, transcripts, references.

③ RAT GENOME DATABASE: The database is for rat species. Display the information about rat's genes, proteins, karyotype, exonic + intronic info, web link to human genome.

④ SACCHAROMYCES GENOME DATABASE: It is a yeast's database. It provides information on gene sequences, their experimental evidences, chromosomal location,

(3) What is an ontology and how does a terminology differ from an ontology? (3 points)

Ontology is a controlled vocabulary working represented in ontology representation language (OKL) which allows for creating modeling objects creation & hence domain models realization with empirical set of grammar rules either strict or relax.

Terminology is not a controlled vocabulary & have no grammar rules. It is more like a definition of a component.

3

(4) What is meant by curation? Provide at least two examples for "curated" databases and describe the essential curation principles that apply to these databases. (4 points)

4

Curation means standardisation and annotation of data so that everyone can understand the process of organising, maintaining, managing & validating a component

Examples:

RCSB PDB - It is a highly curated database. Curator are annotating the submitted protein structure through experimental evidences.

UniProt - Automatic & manual annotation is achieved by TrEMBL & SwissProt respectively they have set of rules & experimental evidences.

MEDLINE - MEDLINE is a highly curated database. Curators curate the abstracts by assigning ~~an~~ indexes from MeSH terms based on the data & info in the ~~doc~~ article.

(5) What are the server-sided programs provided by NCBI to query GEO programmatically? Give an example. (4 points)

The server sided programs provided by NCBI are:-

→ GEO2R ✓

→ BLAST ✓

→ FASTA ✓

Example → The data from GEO is represented as GEO Profiles are taken & then analyzed in R by creating plots & graphs based on data.

(8) What different types of PPI databases are you aware of? Explain briefly each one of them (type of data they contain) and also provide the names of 2 database instances for each type. (5 points) SP

There are 3 types of PPI databases:

(1) Primary Databases: The PPIs are derived from experimental data from large & small researches. It contains data of interactors, interactions, experimental evidences, references
- BioGRID (Biological Interaction of General Repository of Interacting Datasets)
- BIND (Biomolecular Interaction Network Database)

(2) Meta Databases: These databases are created by the integration of primary databases. It provides information the interactors, their reactions along with experimental proofs.
- APID (Agilent Protein Interaction Data analyzer) ✓
- MPID (Microbial Protein Interaction Database) ✓

(3) Predicted Database: It is devised from predicted & experimental data. It is mostly based on experimental evidences but seldom performed in laboratory.
- STRING (Known + predicted PPI) ✓
- MiMIx (Michigan Molecular Interactions) ✓

(9) Explain the term "annotation" and provide at least two examples for databases that are widely known for their high-quality annotations (4 points)

Annotation: Annotation is the collection of comments, references, citations & notations either in a free format or controlled vocabulary to provide extra information related to data. The databases widely known for their annotations are:-

(1) Ensembl
(2) UniProt
(3) RCSB PDB
(4) NCBI-MEDLINE
(5) EMBL-EMA

4

(6) What is the role of MeSH terms and how does MeSH contribute to the computation of "relatedness" of Medline abstracts? (4 points)

MeSH is a controlled vocabulary thesaurus which has a set of naming descriptors that allows retrieval of information at the various levels of specificity.
It allows the indexing for MEDLINE journals + abstracts.
Whenever a search is applied using MeSH terms, it gives the count of MeSH terms that are found in the document. This count can be used by to calculate the relatedness with all the MeSH terms in thesaurus.

3

(7) How do you submit data to ArrayExpress? In what format? Explain. (5 points)

- The data can be submitted to ArrayExpress using Annotare webtool. ✓
- The format is MAGE-ML. ✓
- The data submission needs the following information:-
  - Experiment
  - Protocol
  - Article
  - Molecule Name ??
  - Molecule Description ??

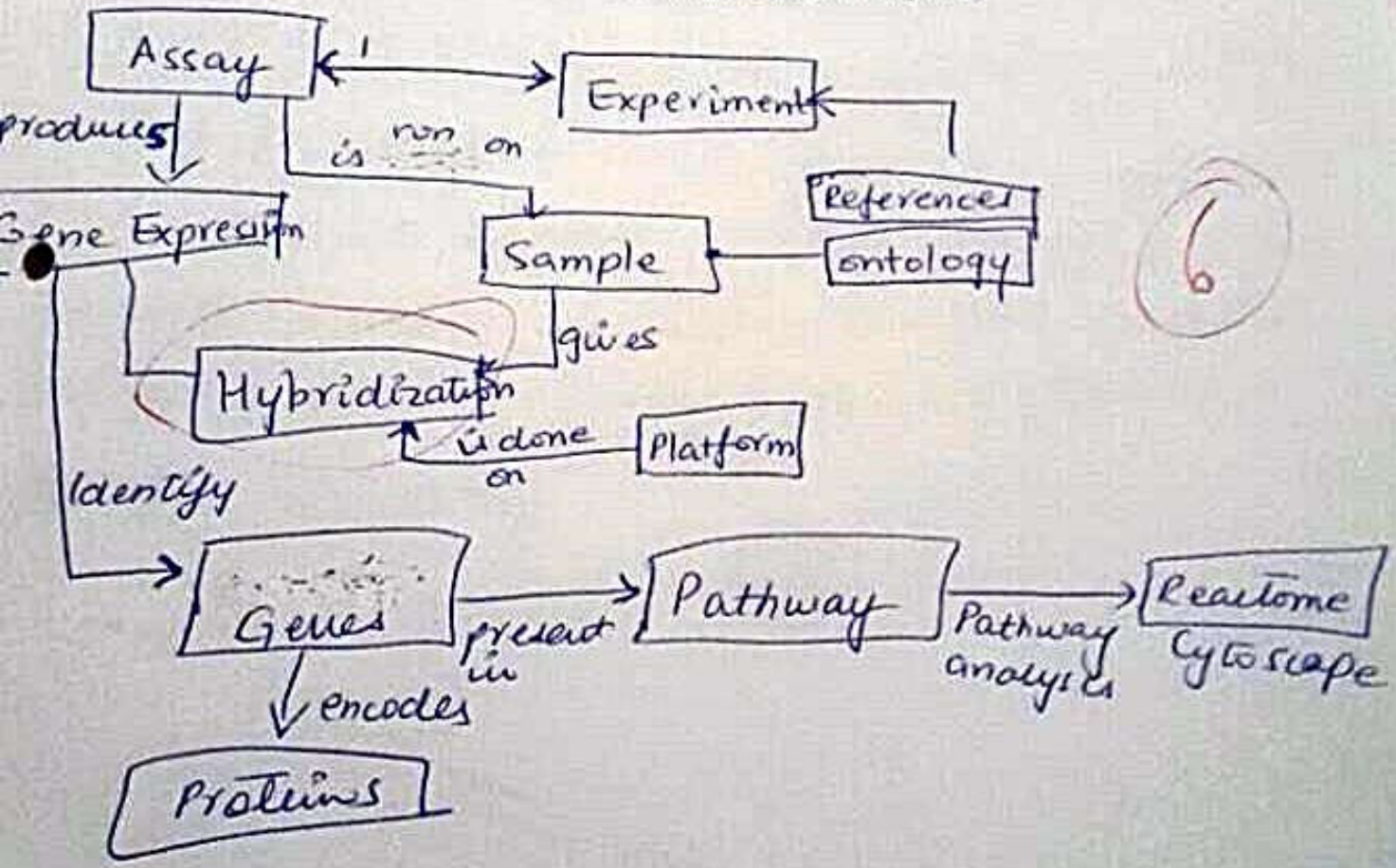
3

(10) What is the difference between Reactome and WikiPathways? (6 points)

Reactome Reactome is highly curated, data containing information of pathways. It is frame based. It is curated by curators only. Pathway analysis is possible. It is not a collaborative, open source. The options opens on the left pane.

WikiPathways: It is collaborative open source for pathways. It can be curated by both scientific community + curators. It is free not frame based. When click on the protein or enzyme, it open in new window.

(11) Develop a conceptual model of a database that links quantitative assay data to pathway information. Sketch that conceptual model as a cartoon; if you can be specific about relationships, draw a entity-relationship diagram. (6 points)



(12) What is UniRef? What are the different types of UniRef databases? (6 points)

UniRef is a part of UniProt. It clusters the proteins based on the similarity between them. There are three parts of UniRef:

① UniRef (100): It clusters those proteins which are having similarity of 100% between them. So they show a strong evolutionary relationships among each other with higher functional similarity.

② UniRef (90): It is clustering much less similarity than UniRef 100 as they show much less seq. similarity & thus clustered in a seperate group.

③ UniRef (50): Sequence similarity is 50% or less than 50%. They are distant homologs, belonging from common ancestor by having a divergent evolution. Less sequence similarity means less structure & function similarity.

(13) How can we validate genetic variants identified in GWAS data? Which other databases are useful for in silico validation of GWAS "signals"? (7 points)

We can validate the genetic variants identified in GWAS by performing another independent study on different cell types or on a different population & by creating a graph & measuring it with each other in order to identify & validate the genetic variants.

The ~~in silico~~ databases that are useful for in silico validation of GWAS signals are:

① Regulome DB:- Provides the DNA sequence features & regulatory proteins using non-coding RNA regions.

② Haploreg DB:- It describes the power of annotation of SNP by providing data on LD ~~plo~~.

③ Regulatory Elements DB: It provides the information about the expression in cellular location & target genes with their scores.

④ Regulome DB GWAS: It provides the information about the SNPs associated with diseases.

4

(14) Which type of information is provided by the following databases: RegulomeDB, HaploReg database and Variant Effect Predictor? (5 points)

RegulomeDB: Provides the DNA sequence features + regulatory proteins using non-coding regions. It provides information about cell targets, genes, GO terms based on an annotation score. The lower the annotation score is the higher the probability is.

HaploReg DB: It provides the information about the power of SNP annotation. It displays the information as r², D, SNPID, genes, chromosomal locations.

Variant Effect Predictor: It is an ENSEMBL based web service. It predicts the mutation effect on a certain mutation in a transcript or gene. It displays the information of PolyPhen scores, SIFT scores, experimental evidences, SNP ids (if related), gene, transcript.

3

(15) How can a DNA mutation affect a protein? Describe some common types of protein level mutation. (4 points)

DNA mutation can affect a protein by altering its folding in 3D structure, it affinity to binding to a certain ligand or protein. & disrupting its intermolecular bonds also.

1. Missense mutation :- change in amino acid.
2. Nonsense mutation :- no change in amino acid.
3. CNV (Copy number variation) :- large insertions + deletions
4. LOH (Loss of heterozygosity) :- loss of a part of chromosomes.
5. Indel : Insertion + deletions (small bases)

2

1,0

(16) Write the python slicing for the DNA sequence seq="ACCTGCTGAA" (1.5 points)

1) first 3 bases

2) last 3 bases

3) bases in position 4 to 6 (included)

1,5P

① print seq[:3] ✓

② Print ~~seq[:3]~~ seq[-3:] ✓

③ print seq[5:6]

because index and position was not clear ✓

(17) What is a web service and describe the general steps how we used it? Which providers of web services do you know? (4 points)

2P

A webservice is an online tool that is distributed by provider + use updated on regular basis + it can be used by the following steps

① Entering the required input.
② Selecting the desired ~~is~~ criteria (or settings to apply on the input. ✓
③ Providing optional input (if available)
④ Running the service.

Providers: NCBI (BLAST) ✓
ENSEMBL (VEP) ✓

(19) What are "abundantly expressed" genes? And what is the distribution of mRNA species in cells? (7 points)

Abundantly expressed genes are those genes which are highly expressed than the normal behaviour of producing more gene products per cell as compared to the normal cell.

There are 2 types of distribution-

Qualitative Distribution: This distribution is cell specific & is by the reg production of regulatory mRNAs.

Quantitative Distribution: This distribution is based on abundantly expressed genes but not on the regulatory genes.

(2)

20) You are doing your Master Thesis with some clinical researcher at Veruniberg. The medical researchers expect you to help them with in silico methods to identify molecular determinants of Colorectal Carcinoma. What databases will you access and mine to support them and what sort of information do you get from these databases? (7 points)

(7)

Following databases will be accessed:

(1) ENSEMBL (to get the info about genetic variants + genes).

(2) Omim (to ~~pro~~ get the info about inherited variants + their etiology).

(2) SNP (to assess the related SNP info about the molecule).

(4) RegulomeDB GWAS (to have the information about SNP associated with disease).

(5) UniProt (to have the protein specific information).

(6) Pathway DB - KEGG (to find the process that are inhibited or activated by the molecule or the location in pathway).

(7) GEO (to have the assess for microarray expression of the related mutation)