# Midterm Exam

**READ ME FIRST**

For the multiple choice, pick one and only one best answer. For multiple choice questions, if you don't like any of your options you may write one complete sentence *explaining your logic* below the set of provided answers and I will consider it for partial credit. Without the explanation I will not consider alternative answers.

Short answer questions provide the required information.

112 points possible.

**Part 1: Theory (80)**

1. (4) What is meta-data?
   a. The data dictionary maintained by the DBMS
   b. Data describing the data
   c. The semantic domain of a model
   d. An ontology developed for a model.

2. (4) If a file system is used for managing data then applications must be provided with a location in a file to obtain the needed information. This is called
   a. Structural dependence
   b. Access dependence
   c. Logical independence
   d. Hardware independence

3. (8) List the 8 Fundamental operators of relational algebra and their mathematical symbols.

| | |
|---|---|
| σ Selection | ρ Rename |
| Π Projection | \|X\| Join |
| X Cartesian Product | - Difference |
| ∪ Union | / Divide |

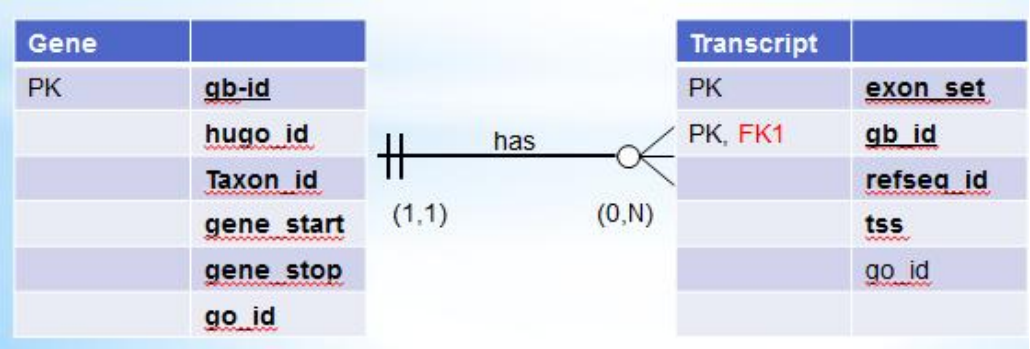
## Midterm Exam

| TAXON | |
|---|---|
| A | TAXON_ID |
| B | Name |

| GENE | |
|---|---|
| C | Gene_ID |
| D | Gene_Name |
| E | Taxon_ID |
| F | G_Start1, G_Start2, G_Start3 |
| G | G_Stop |
| H | G_Length1, G_Length2, G_Length3 |

| mRNA | |
|---|---|
| I | Transcript_ID |
| J | Gene_ID |
| K | Transcript_Start |
| L | Transcript_Stop |
| M | Transcript_Length |

4. (4) In the above figure
   a. Which rows define the primary keys? A, C, I, J
   b. Which rows are Not Null? A, B, C, D, I, J, K, L
   c. Which rows contain multi-valued attributes? F, H
   d. Which rows indicate a potential foreign key relationship? A to E and C to J




5. (4) Relationship strength depends on how the identifying attribute is formulated and relationship participation depends on how the process is defined. In the above figure we have:
   a. Mandatory participation and a strong transcript entity
   b. Strong relationship and strong transcript entity
   c. Mandatory participation by the transcript entity and a strong relationship
   d. Optional participation by the transcript entity and a weak entity

6. (4) A SAM file header produced on an Illumina platform includes a sequence identifier for the read within the set, an instrument identifier and run date and information as to whether this was a single or paired-end run. This is an example of what type of attribute?

   a. A simple attribute
   b. A composite attribute
   c. A single-valued attribute
   d. A multi-valued attribute

7. (4) Which of the following describes a recursive relationship with a unary degree?

a. An Enzyme entity with a relationship to an Enzyme entity where the relationship description is 'modifies enzyme active site'
b. A Microarray Probe entity with a relationship a Target entity 'probe binds to target'.
c. An enzyme entity with a relationship to an Enzyme-Substrate bridge entity 'converts to product'.
d. An Enzyme entity with a relationship to a Substrate entity 'converts' and a relationship to a Chemical entity 'produces'.

8. (4) When you progress from the conceptual model to a model whose logic consists of set operations you have to modify the representation of many-to-many relationships. For the relational model this requires

   a. Redefine optional relationships as mandatory and replicate attributes in child tables
   b. Decompose their composite attributes into additional atomic attributes
   c. Decompose to binary 1:M relationships via linking tables
   d. Make sure that only binary relationships are present

9. (4) Most RDBMS allow the database owner to drop a foreign key constraint in order to bulk-load data very quickly. Which of Codd's Rules does this subvert?

   a. View updating must be possible when new data is inserted into entities.
   b. You must be able to insert, delete and modify data using set-level functions
   c. You cannot bypass the integrity rules, whatever your access privileges.
   d. Integrity constraints must be stored in the database system catalog.

10. (4) I have created a logical model for a project that consists of 24 tables, including the necessary linking tables. I use MySQL Workbench to convert my design to SQL statements, export those and run the command line of SQLite. When I look at the database system summary I find that there are 30 tables. What went wrong?

    a. Not all of the many-to-many relations were expressed properly and the system has attempted to optimize the model for you.
    b. The system must be self-describing so additional tables have been created to manage the meta-data.
    c. Several of the attributes were multi-valued and the system has moved them to another table.
    d. There were several ternary relationships and the system has decomposed them into binary sets of relationships.

11. (4) What are the operations used to combine conditions on sets to create complex expressions?

a. The Set operators: Union, Intersection, and Difference.
b. The Logical operators: AND, OR, NOT and IDENTITY
c. The Algebraic operators: +, -, *, ÷, exp, log
d. The Comparison operators: +, -, =, <, >, ≠, ≤, ≥

12. (4) A first-order predicate expression is evaluated on conditions by what criteria?

a. Whether it equals a specified value
b. Whether every value in the tuple equals the referenced tuple values.
c. Whether the restrictions on the table combine selection and projection.
d. Whether the condition is true or false for the operations specified.

## Midterm Exam

GENE_SNP

| Gene_Name | Locus | SNP_ID | AssocScore | ExpFoldChange |
|---|---|---|---|---|
| TNFRSF14 | 1p36 | rs3890745 | 3.6E-6 | 0.9 |
| PTPN22 | 1p13 | rs2476601 | 9.1E-74 | 1.1 |
| REL | 2p16 | rs13031237 | 7.9E-7 | 1.2 |
| AFF3 | 2q11 | rs10865035 | 2.0E-6 | 2.1 |
| STAT4 | 2q32 | rs7574865 | 2.9E-7 | 0.8 |
| CTLA4 | 2q33 | rs3087243 | 1.2E-8 | 1.0 |
| HLA-DRB1 | 6p21 | rs6910071 | 1.0E-299 | 1.1 |
| TNFAIP3 | 6q23 | rs6920220 | 8.9E-13 | 3.4 |
| TRAF1_C5 | 9q33 | rs5029937 | 7.5E-8 | 2.5 |
| PRKCQ | 10p15 | rs3761847 | 2.1E-7 | 0.8 |
| CD40 | 20q13 | rs4750316 | 2.8E-9 | 1.0 |
| RRAD | 16q22 | rs368384610 | 2.2E-11 | 2.4 |

13. (4) For the table shown above, write a *relational algebra expression* to select for the rows in GENE_SNP with a locus on the second chromosome that also have an expression fold change that is greater than or equal to 1.0. User the correct symbol notation. Show the resulting relation as a table.

$\sigma_{Locus=2***\ ,\ ExpFoldChange \geq 1.0}$ (GENE_SNP)

Gene_SNP

| Gene_Name | Locus | SNP_ID | AssocScore | ExpFoldChange |
|---|---|---|---|---|
| REL | 2p16 | rs13031237 | 7.9E-7 | 1.2 |
| AFF3 | 2q11 | Rs10865035 | 2.0E-6 | 2.1 |
| CTLA4 | 2q33 | Rs3087243 | 1.3E-8 | 1.0 |

14. (4) Project on the resulting relation above for the Gene_Name, AssocScore and SNP_ID columns, and rename the resulting table SNPscore. Write the *complete algebraic expression* and show the final table.

$\rho_{ASSOC\_SNP}$ ($\Pi_{Gene\_Name,AssocScore,SNP\_ID}$ ($\sigma_{Locus=2***\ ,\ ExpFoldChange \geq 1.0}$ (GENE_SNP))

ASSOC_SNP

| *Gene_Name* | AssocScore | SNP_ID |
|---|---|---|
| *REL* | 7.9E-7 | rs13031237 |
| *AFF3* | 2.0E-6 | Rs10865035 |
| *CTLA4* | 1.3E-8 | Rs3087243 |

**Midterm Exam**

GENE

| Gene_Symbol | (Length_mRNA) | Location |
|---|---|---|
| UB148 | 1358 | Golgi |
| SLC24 | 1590 | ER |
| ALTN | 2001 | Nucleus |
| RBC2 | 908 | ER |

X

PROTEIN

| Protein_id | Gene_Symbol | Primary_Fn |
|---|---|---|
| 139002 | UB148 | degradation |
| 145923 | SLC24 | signaling |
| 209974 | ALTN | skeleton |
| 255601 | RBC2 | translation |

15. (4) If you carry out the Product operation on the two tables above,
    a. What will the header look like (the list used to define a table)?

(GENE X PROTEIN) →
(GENE.Gene_symbol, GENE.Length_mRNA, GENE.Location, PROTEIN.Protein_id, PROTEIN.Gene_Symbol, PROTEIN. Primary_Fn)

b. How many rows will you have?
$2^4 = 16$

**Midterm Exam**

GENE

| Gene_Symbol | (Length_mRNA) | Location |
|---|---|---|
| UB148 | 1358 | Golgi |
| SLC24 | 1590 | ER |
| ALTN | 2001 | Nucleus |
| RBC2 | 908 | ER |

X

PROTEIN

| Protein_id | Gene_Symbol | Primary_Fn |
|---|---|---|
| 139002 | UB148 | degradation |
| 145923 | SLC24 | signaling |
| 209974 | ALTN | skeleton |
| 255601 | RBC2 | translation |

16. (8) Using the two tables above, show the results of the Natural Join, Theta Join (restrict to having a first letter between M and Z) and EquiJoin.

Natural Join (values must match in all attributes with the same name):

| Gene.Gene_Symbol | GENE.Length_mRNA | GENE.Location | PROTEIN.Protein_id | PROTEIN.Primary_Fn |
|---|---|---|---|---|
| UB148 | 1358 | Golgi | 139002 | degradation |
| SLC24 | 1590 | ER | 145923 | signaling |
| ALTN | 2001 | Nucleus | 209974 | skeleton |
| RBC2 | 908 | ER | 255601 | translation |

Theta (duplicate column remains) – condition means ALTN row is gone and there is a PROTEIN.Gene_Symbol column

Equi join - four rows returned.

**Midterm Exam**

GENE

| Gene_Symbol | (Length_mRNA) | Location |
|---|---|---|
| UB148 | 1358 | Golgi |
| SLC24 | 1590 | ER |
| ALTN | 2001 | Nucleus |
| RBC2 | 908 | ER |

X

PROTEIN

| Protein_id | Gene_Symbol | Primary_Fn |
|---|---|---|
| 139002 | UB148 | degradation |
| 145923 | SLC24 | signaling |
| 255601 | RBC2 | translation |

17. (4) Using the two tables above, show the results of the Left Outer, Right Outer and Full Join operations.

Left Outer Join:

| Gene.Gene_Symbol | GENE.Length_mRNA | GENE.Location | PROTEIN.Protein_id | PROTEIN.Primary_Fn |
|---|---|---|---|---|
| UB148 | 1358 | Golgi | 139002 | degradation |
| SLC24 | 1590 | ER | 145923 | signaling |
| ALTN | 2001 | Nucleus | null | null |
| RBC2 | 908 | ER | 255601 | translation |

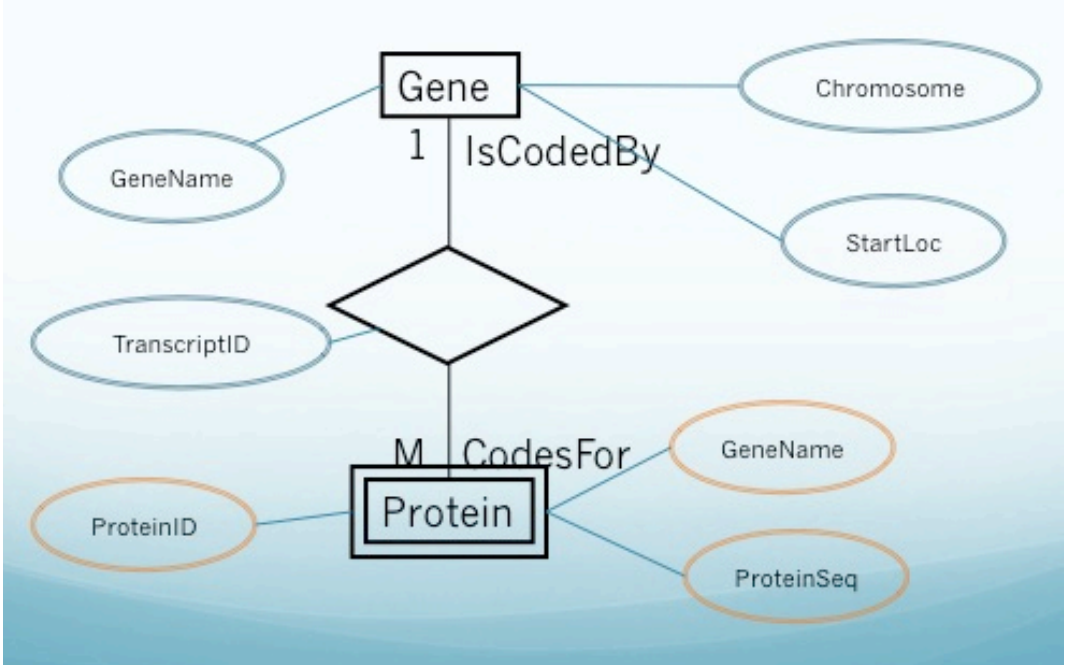
Right outer –

Full -

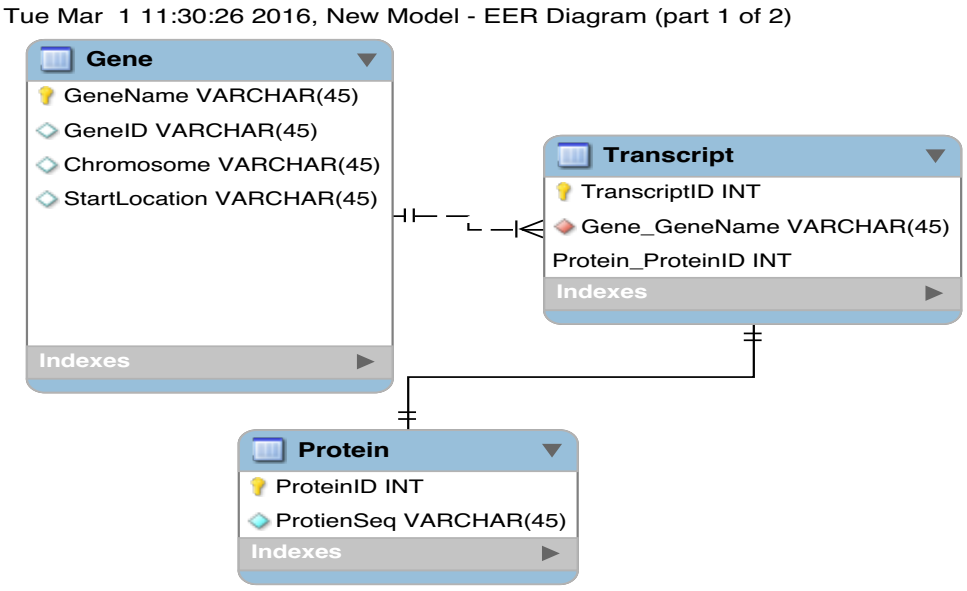
**Midterm Exam**

**Part2: Diagrams and SQL (32)**

Figure 1: Chen Diagram

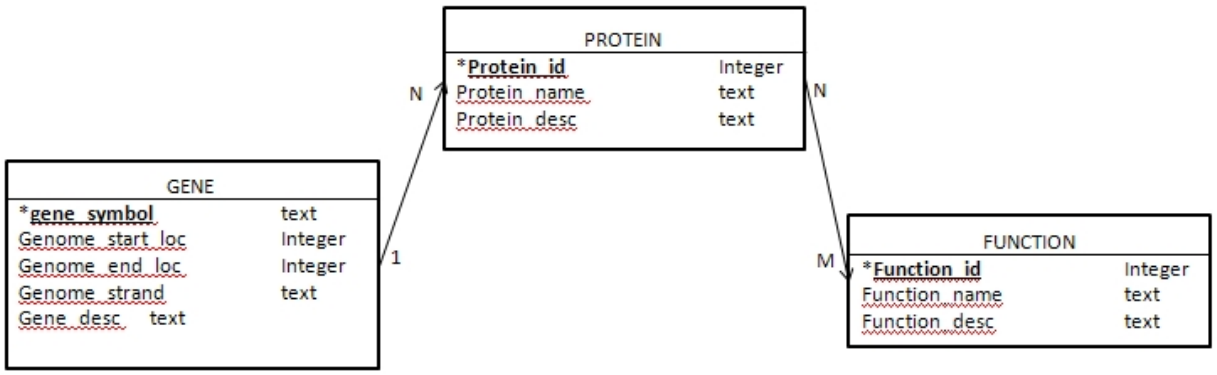



1. (8)Generate the ER diagram from the Chen Diagram above. (Figure 1.)



**Midterm Exam**

Figure 2: GENE/PROTEIN/FUNCTION ERD




For these two problems use the ERD shown in Fig. 2

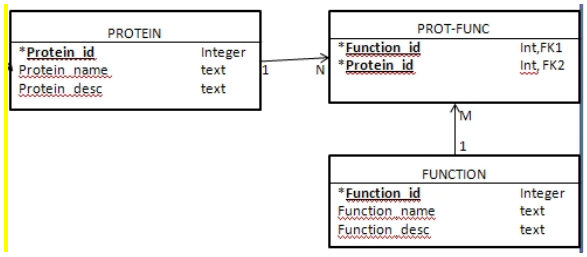
2. (8) Write the SQL to create the GENE and PROTEIN tables shown in Fig. 2. Make sure to include:
   - Any additional attribute(s) necessary to implement the relationship between the two tables.
   - Data types and table constraints, such as uniqueness and empty set limits.

```
CREATE TABLE GENE
        (
                gene_symbol NOT NULL UNIQUE text,
                genome_start_loc integer,
                genome_end_loc integer,
                genome_strand text,
                gene_desc text,
                PRIMARY KEY(gene_symbol),
                UNIQUE (genome_start_loc,  genome_end_loc,  gene_desc)
        );

        CREATE TABLE PROTEIN
        (
                protein_id  NOT NULL UNIQUE integer,
                protein_name text,
                protein_desc text,
                gene_symbol text,
                PRIMARY KEY(protein_id),
                FOREIGN KEY(gene_symbol) REFERENCES GENE
        );
```

**Midterm Exam**

3. (8) How would you handle the design problem presented by the many-to-many relationship between the PROTEIN and FUNCTION entities so that it can be implemented in the relational database? Show the diagram, the SQL to implement it and explain your reasoning with a sentence.




You would need to implement a linking table, PROT_FUNC that takes uses the PK of each parent as part of its PK, with a FK to establish the relation.

```
CREATE TABLE PROT_FUNC
        (
                Protein_id      NOT NULL UNIQUE integer,
                Function_id     NOT NULL UNIQUE integer,
                PRIMARY KEY (Protein_id, Function_id),
                FOREIGN KEY (Protein_id) REFERENCES PROTEIN
                FOREIGN KEY (Function_id) REFERENCES FUNCTION
);
```

## Midterm Exam

Table 1: GeneOperonLength

| operon_id | gene_id | gene_length |
|---|---|---|
| 2 | deoA | 1322 |
| 2 | deoB | 1223 |
| 1 | mxaA | 989 |
| 2 | deoC | 779 |
| 2 | deoD | 719 |
| 3 | flgA | 659 |
| 1 | moaB | 512 |
| 1 | moaC | 485 |
| 1 | moaE | 452 |
| 3 | flgN | 416 |
| 3 | flgM | 293 |
| 1 | moaD | 345 |

4. (4) Using the table above write the SQL to retrieve all of the operon_id and gene_id where the gene length is less than the average.

```
SELECT operon_id, gene_id,
FROM GENEOPERONLENGTH
WHERE gene_length<(SELECT AVG (gene_length) FROM GENEOPERONLENGTH);
```

5. (4) Using Table 1 write the SQL to retrieve the number of gene_id containing ‘m’ and ‘a’.

```
SELECT  COUNT (gene_id),
FROM GENEOPERONLENGTH
WHERE gene_id LIKE ‘%m%a’;
```