**CSE 601 – FALL 2016**

**DATA MINING AND BIO INFORMATICS**

**PROJECT 1 – PART 1 – REPORT**

**Submitted By:**

Arti Gupta (50170010)

artigupt@buffalo.edu

Sudharchith Sonty (50169912)

[sudharch@buffalo.edu](mailto:sudharch@buffalo.edu)

**Introduction:**

In this assignment, we have considered clinical and genomic data for biomedical data warehousing and mining. We have designed a logical data model using the given sample data tables and their attributes. Using the concept of fact constellation of star schemas, analysis of such a complex dataset has been explained.

**Datasets:**

Various tables have been used as per the given dataset like sample, patient, experiment, drugs, disease, clinical\_fact etc. and their relationship have been provided.

**PART 1: CREATING WAREHOUSE AND LOADING DATA:**

We have designed a fact constellation of several star-schemas. These are:

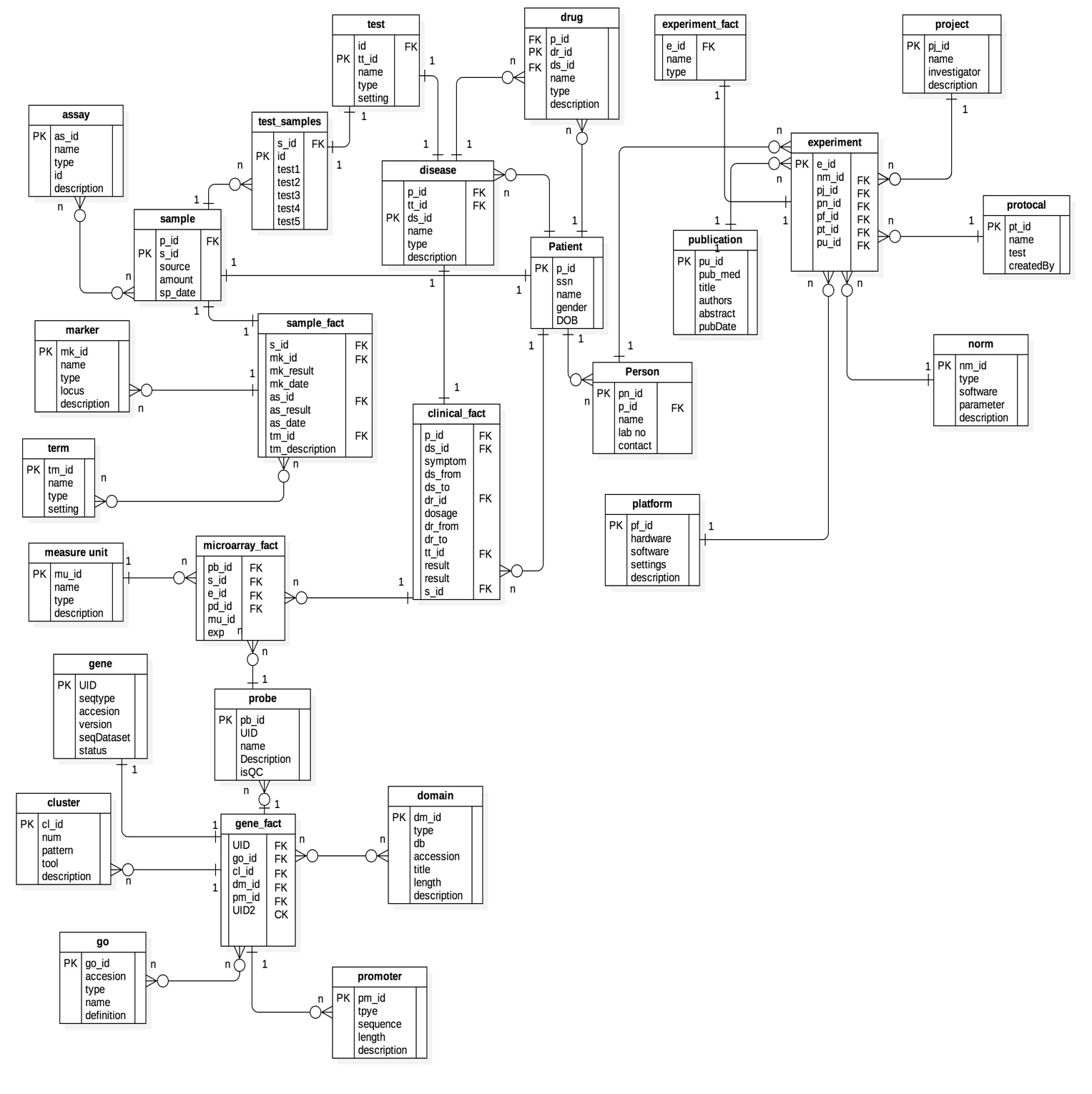
1. Patient is the fact entity of our schema. p\_id is its primary key. It has been connected with dimension drugs, disease, clinical\_fact, person and sample.
2. Patient is the core table for this biological data. One patient can have many diseases. So, it will be one-to-many relationship.
3. One patient can use various drugs for a disease so, it will be one to many relationship of drugs with both patient and disease table.
4. Every patient will be linked to person entity which stores name, contact no etc. information of each patient.
5. One sample can be used for a patient at a time, so it will be one-to-one relationship where p\_id will be used as foreign key in sample entity.
6. Clinical\_fact entity is connected to patient, disease and microarray\_fact entity and stores foreign keys.

Thus, patient table is the fact entity and other entities are its dimensions.

1. Sample entity is the fact for another star schema. s\_id is the primary key. Which is also linked with patient fact. This entity has its own dimension entities like assay, test\_samples, sample\_fact.
2. One sample can be taken from a patient at a time. These samples can be used as any number of samples for various tests. So, it will be one-to-many relationship with test\_samples.
3. On the basis of amount of samples, no. of tests can be performed to diagnosis a disease i.e. one test sample is used to test one disease. Therefore, test\_samples will be one-to-one relationship with test.
4. Again, one test will identify for one disease so it will be one-to-one relationship between test entity and disease entity. Alongside, it is linked with drugs entity, thus storing a relationship with Patient fact entity.
5. Now, sample entity is also connected to sample fact entity that stores the primary keys of marker entity, term entity as foreign keys in it.

Thus, Sample entity is the fact entity and adding dimensions to it, further making a star schema.

**Logical Data Model:**



**Fig: Schema design for clinical and gene data for biomedical data warehousing**

**RESULTS FOR SAMPLE QUERIES**

1. **QUESTION:** List the number of patients who had “ tumor” (disease description), “ leukemia” (disease type) and “ ALL” (disease name), separately.

**QUERY:**

*select patient\_count\_ALL, patient\_count\_leukemia, patient\_count\_tumor from (select count(distinct c.p\_id) patient\_count\_ALL from clinical\_fact c, disease d where d.name='ALL' and c.ds\_id=to\_char(d.ds\_id)) a, (select count(distinct c.p\_id) patient\_count\_leukemia from clinical\_fact c, disease d where d.type='leukemia' and c.ds\_id=to\_char(d.ds\_id)) b, (select count(distinct c.p\_id) patient\_count\_tumor from clinical\_fact c, disease d where d.description='tumor' and c.ds\_id=to\_char(d.ds\_id)) c*

**RESULT:**

1. **QUESTION:** List the types of drugs which have been applied to patients with “ tumor”.

**QUERY:**

*"select distinct a.type*

*from drug a,*

*(select distinct c.p\_id, c.dr\_id*

*from clinical\_fact c, disease d*

*where d.description='tumor'*

*and c.ds\_id=to\_char(d.ds\_id)) b*

*where to\_char(a.dr\_id)=b.dr\_id",*

**RESULT:**

1. **QUESTION:** For each sample of patients with “ ALL”, list the mRNA values (expression) of probes in cluster id “ 00002” for each experiment with measure unit id = “ 001”. (Note: measure unit id corresponds to mu\_id in microarray\_fact.txt, cluster id corresponds to cl\_id in gene\_fact.txt, mRNA expression value corresponds to exp in microarray\_fact.txt, UID in probe.txt is a foreign key referring to gene\_fact.txt)

**QUERY:** *"select distinct s\_id, exp from microarray\_fact where*

*pb\_id in (select pb\_id from probe p where u\_id in*

*(select u\_id from gene\_fact g where cl\_id='2')) and mu\_id='1'*

*and s\_id in (select distinct s\_id from clinical\_fact cf where p\_id in (*

*select distinct p\_id from clinical\_fact c, disease d*

*where d.name='ALL' and c.ds\_id=to\_char(d.ds\_id)) and s\_id != 'null')")*

**RESULT:**

1. **QUESTION:** For probes belonging to GO with id = “ 0012502”, calculate the t statistics of the expression values between patients with “ ALL” and patients without “ ALL”. (Note: Assume the expression values of patients in both groups have equal variance, use the t test for unequal sample size, equal variance)

**QUERY:**

**RESULT:**

1. **QUESTION:** For probes belonging to GO with id=“ 0007154”, calculate the F statistics of the expression values among patients with “ ALL”, “ AML”, “colon tumor” and “ breast tumor”. (Note: Assume the variances of expression values of all four patient groups are equal.)

**QUERY:**

**RESULT:**

1. **QUESTION:** For probes belonging to GO with id=“ 0007154”, calculate the average correlation of the expression values between two patients with “ ALL”, and calculate the average correlation of the expression values between one “ ALL” patient and one “ AML” patient. (Note: For each patient, there is a list of gene expression values belonging to GO with id=“ 0007154”. Suppose you get 𝑁1 “ ALL” patients and 𝑁2 “ AML” patient. For the average correlation of the expression values between two patients with “ ALL”, you need first calculate 𝑁1 × (𝑁1 − 1)/2 Person Correlations then calculate the average value. For the average correlation of the expression values between one “ ALL” patient and one “ AML” patient, you need first calculate 𝑁1 × 𝑁2 Person Correlations then calculate the average value.)

**QUERY:**

**RESULT:**

**PART-3: OLAP OPERATIONS AND KNOWLEDGE DISCOVERY**

1. **QUESTION:** Find all the patients with “ALL” (group A), while the other patients serve as the control (group B).

* **QUERY:**
* **RESULT:**

1. **QUESTION:** For each gene, calculate the t-statistics for the expression values between group A and group B.

* **QUERY:**
* **RESULT:**

1. **QUESTION:** If the p-value of the t-test is smaller than 0.01, this gene is regarded as an “informative” gene.

* **QUERY:**
* **RESULT:**