**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:

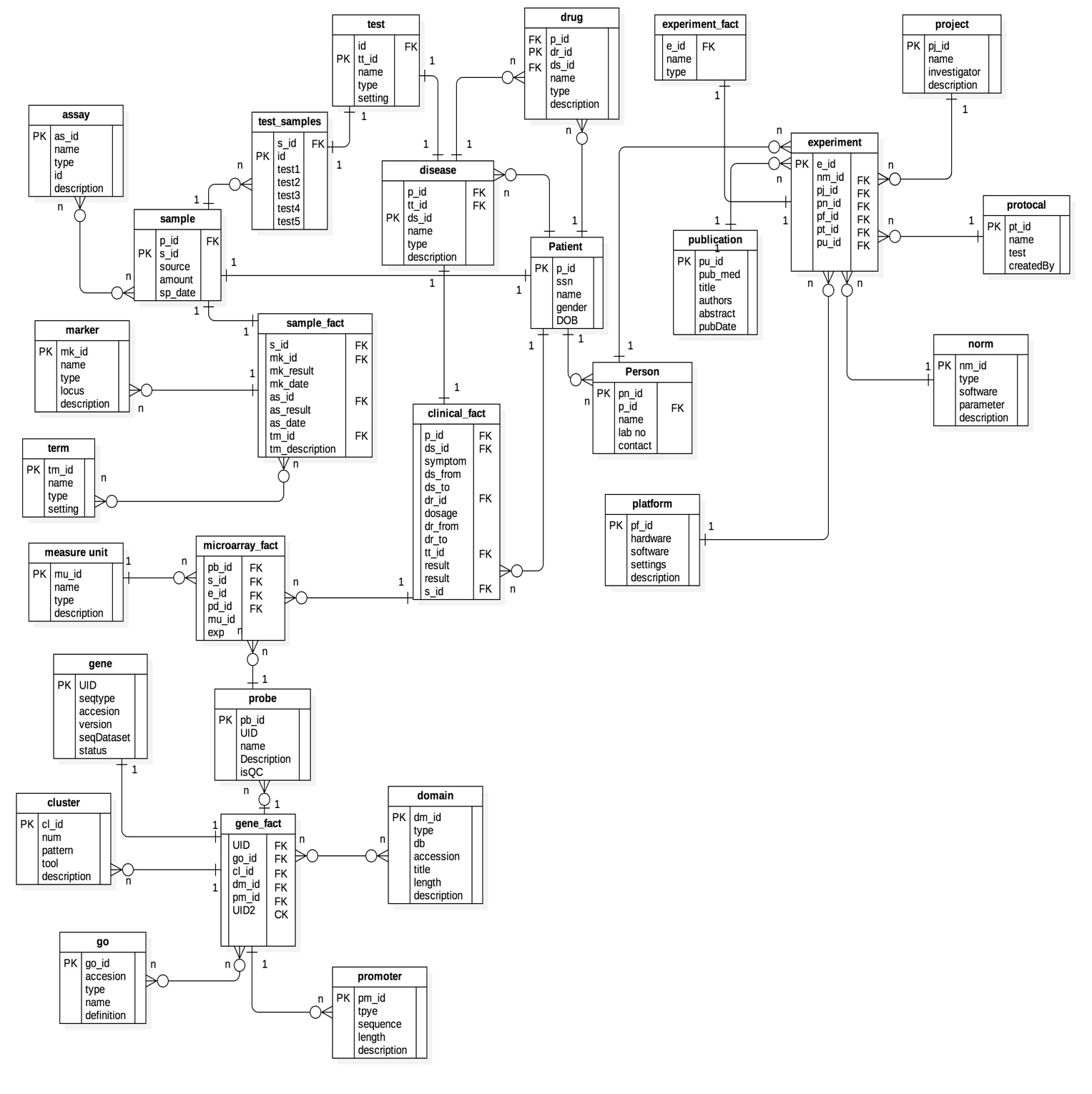
1. Patient is the fact entity of our schema. p\_id is its primary key. It has been connected with dimension drugs, disease, clinical\_fact, person and sample.
2. Patient is the core table for this biological data. One patient can have many diseases. So, it will be one-to-many relationship.
3. One patient 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 “ tumour” (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:**

PATIENT\_COUNT\_ALL PATIENT\_COUNT\_LEUKEMIA PATIENT\_COUNT\_TUMOR

---------------------- ---------------------- ----------------------

13 27 53

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 (SAMPLE):**

TYPE

--------------------------------------------------

Drug Type 015

Drug Type 011

Drug Type 018

Drug Type 019

Drug Type 004

Drug Type 005

Drug Type 003

Drug Type 006

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 (SAMPLE):**

S\_ID EXP

-------------------- --------------------

973218 42

973218 13

580723 11

580723 186

580723 73

953971 8

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)

* **RESULT:** Welch Two Sample t-test

data:  e1 and e2

t = -0.98909, df = 512.58, p-value = 0.3231

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.551743   3.815205

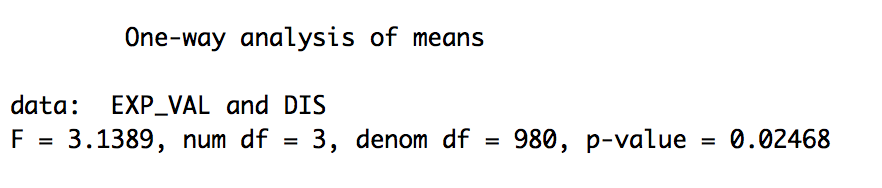
sample estimates:

mean of x mean of y

95.93590  99.80417

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.)

**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).

* **RESULT:** Created two groups in R.

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

* **RESULT (SAMPLE):**

GENE D\_ALL D\_OTHER T\_STATS P\_VALUE

------------ ----------- ------------ ----------- ------------

198293 86.85 99.5 -0.89 0.38

397177 93.42 88 0.45 0.66

498845 125.46 103.51 1.78 0.08

794927 91.69 108.42 -1.35 0.18

822095 93.42 101.2 -0.56 0.57

843355 129.77 98.11 2.33 0.02

920984 96.58 115.86 -1.51 0.14

1030083 103.46 98.17 0.4 0.69

1059822 68 97.03 -2.21 0.03

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

* **RESULT(SAMPLE):**

GENE

----------------------

1305056

1433276

4826120

7396036

10295571

10595583

11333636

11900023

13947282

15295292

16073088

17144710

17791557

18493181

18634953

21633757

**PART 3.2 – OLAP OPERATIONS AND KNOWLEDGE DISCOVERY**

1. **QUESTION:** Find the informative genes w.r.t. “ALL”

**RESULT(SAMPLE):**

U\_ID

----------------------

1305056

10595583

17144710

21633757

23552119

38558690

53478188

62515965

62908060

72920004

74496827

75434512

78742292

1. **QUESTION:** Find all the patients with “ALL” (group A).

**RESULT(SAMPLE):** Created a VIEW

P\_ID

--------------------

6060

33553

79175

765

22162

47360

70863

65736

13258

77689

47880

2378

58484

1. **QUESTION:** For each patient PA in group A, calculate the correlation rA of the expression values of the informative genes between PN and PA.

**RESULT(SAMPLE):**

P\_ID RA

---------------------- ----------------------

765 0

2378 0

6060 0

13258 0

22162 0

33553 0

47360 0

47880 0

58484 0.00000000000000011103441356384342

65736 0

70863 0

77689 0.0000000000000001110344421632972

79175 0.00000000000000011103441356384342

1. **QUESTION:** Patients without “ALL” serve as the control (group B).

**RESULT(SAMPLE):**

P\_ID

--------------------

68707

28272

70045

70835

56425

75733

6413

17659

* **QUESTION:** For each patient PB in group B, calculate the correlation rB of the expression values of the informative genes between PN and PB.

**RESULT(SAMPLE):**

P\_ID RB

---------------------- ----------------------

304 0

2318 0

6413 0

13243 0

14087 0.00000000000000011103412212660645

17659 0

21772 0

22176 0

28272 0.00000000000000011103385661814893

28582 0

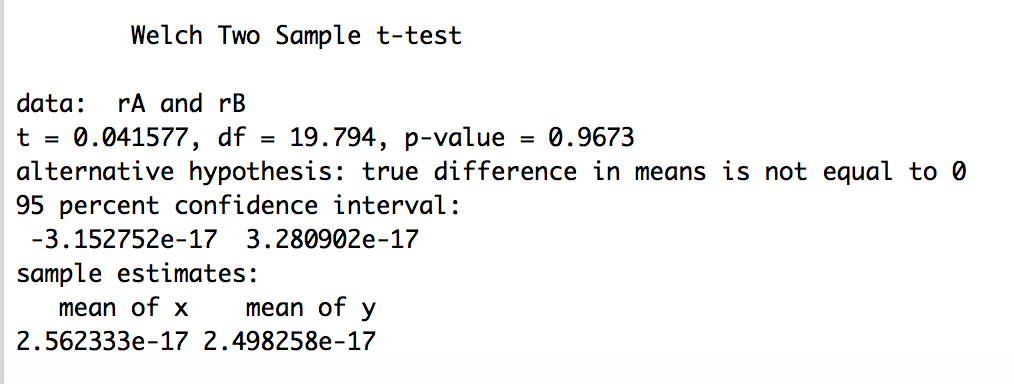
31076 0

34267 0

43487 0.00000000000000011103385661814893

1. **QUESTION:** Apply t-test on rA and rB, if the p-value is smaller than 0.01, the patient is classified as “ALL”

**RESULT(SAMPLE):**

****

**INFORMATIVE\_GENE\_ALL**

**----------------------**

**65736**

**47880**

**77689**

**58484**

**79175**

**33553**

**47360**

**6060**

**2378**

**13258**

**INFORMATIVE\_GENE\_NOTALL**

**-----------------------**

**45652**

**44019**

**17659**

**31076**

**70045**

**13243**

**68309**

**28582**

**70835**

**79352**

**92978**