20.453J / 2.771J / HST.958J Biomedical Information Technology
Fall 2008

DESIGN AND IMPLEMENTATION OF BIOLOGICAL PATHWAY INTERACTION DATABASE SYSTEM (PID)

1. BACKGROUND

1.1 *Objectives*

Upon completion of the assignment, the student should be able to:

- Construct an enhanced entity relationship model at a conceptual level
- Map the model into a relational database system
- Implement the given schema on a relational DBMS
- Use a database language for manipulating and querying the data

1.2 *Nature of Assignment*

This is an individual assignment. Discussion is allowed. However, submissions must be your original work. Plagiarism will be heavily penalized.

1.3 *Hardware and Software Resources*

Machines	DBMS
Participant's laptop	PostgreSQL server

1.4 *Introduction*

The assignment covers the portion of the course concerning data modelling, database design and implementation. However, you are encouraged to include aspects for improving the efficiency, presentation etc.

The overall aim of the assignment is to develop an application based on a given data model using a chosen database management system. This exercise will bring you through the crucial first part of the life cycle of a database application. The PID database will be used as a case study. The assignment is made up of 2 components (Part A and Part B).

Part A involves:

- constructing the Entity-Relationship (ER) model for PID

Part B involves:

- mapping the ER model to a relational database schema
- implementing the database using PostgreSQL. Data for populating the database are adapted from the Syndecan-1-mediated signaling events pathway in PID.
- constructing SQL query statements

2. DESCRIPTION OF THE ASSIGNMENT

The description of the application is given in the appendices. This includes the background and general requirements of the application, conceptual information about the system and a list of queries that must be fulfilled as a minimum.

All submissions must be type-written. That is, no hand written solution will be accepted.

2.1 *Part A: Creating an ER Diagram*

The appendices give conceptual information about the database requirements. Based on the provided information, construct a suitable ER diagram.

The followings are required to be submitted to the stellar website by **September 30th, 2008 (12 AM Singapore Time)**.

- ER Diagram
- Semantic description of the attributes, entity-sets and relationships in your diagram
- Any assumptions made for your design
- Indicate all the feature(s)/specification(s) (if any) that CANNOT be captured by ER-diagram.

Submission must be named in the following convention: ***<FamilyName>_HW1A.<FileFormat>***

2.2 *Part B: Mapping of the ER Diagram onto Relations and Implementation of the Schema*

The ER diagram (your solution) is to be mapped onto relations (tables). Follow the general guidelines covered during the lectures to produce suitable tables. The solution must clearly state the primary and foreign keys, the data type, the integrity constraints, additional constraints etc. It is expected that you submit your definition in textual form, e.g.

```
CREATE TABLE name (
        attr1 Datatype NOT NULL,
        attr2 Datatype,
        …
        PRIMARY KEY (attr1),
        FOREIGN KEY (attr3) REFERENCES name(attr1)
                ON DELETE … ON UPDATE …,
);
```

A graphical representation of your solution can be submitted in addition, however, this cannot substitute the textual schema definition.

Solve the laboratory task by using the PostgreSQL Server – no other programs and submission types are allowed. The ***PostgreSQL_Mini_Installation_User_Guide.doc*** provides you with additional information.

The schema has to be implemented in the provided relational DBMS, i.e. queries, and constraint have to be created. Make full use of the DBMS's features to implement data integrity requirements (such as primary, foreign keys and unique constraints). Your database should be populated with data (***Data_Adapted_From_PID.xml***) adapted from the Syndecan-1-mediated signalling events pathway in Pathway Interaction Database (PID) for the demonstration of the DBMS with respect to the queries, constraints etc.

Course participants in Singapore are expected to demonstrate your project. Information regarding the demonstration will be provided at a later date. During the demo session, it is expected that:

- New queries can be written and produce satisfying results.
- Questions regarding the design and related issues can be answered appropriately.
- The proper working of the DBMS is demonstrated.

Course participants in MIT are expected to submit the database implementation file (.backup) and the query implementation file (.sql) for the following queries:

- **Query 1:** Count the number of *interactions* involved in "Syndecan-1-mediated signaling events" pathway (Pathway ID: 200036)
- **Query 2:** List the *molecule_ID*, *name*, *EntrezGene_ID* and *UniProt_ID* of the proteins with aliase "SYND1"
- **Query 3:** List the *names* and *molecule_ID* of complexes having protein with UniProt_ID "O14936" as its component.
- **Query 4:** List all the *references* providing physical interaction inference support (Evidence: IPI) for "Syndecan-1-mediated signaling events" pathway (Pathway ID: 200036). Arrange the output in ascending order of reference_ID.
- **Query 5:** For pathway "Syndecan-1-mediated signaling events" pathway (Pathway ID: 200036), list the *name*, *molecule_ID* of all proteins involved in "modifications" interaction and state the *interaction_ID* and *role* of these proteins in these modifications interactions. Arrange the output in ascending order of molecule_ID followed by interaction_ID.

The followings are required to be submitted to the class website by **October 9th, 2008 (12 AM Singapore Time)**:

All participants (Singapore and MIT)

- Textual form of relational schema
- Semantic description of the attributes and relations
- Assumptions (if any)

MIT participants only

- Database implementation in .backup file format
- SQL queries implementation in .sql file format

Submissions must be named in the following convention: ***<FamilyName>_HW1B.<FileFormat>***

APPENDIX A: CONCEPTUAL INFORMATION ABOUT PID

The Pathway Interaction Database (PID, http://pid.nci.nih.gov) [1, 2], created by U.S. National Cancer Institute and Nature Publishing Group contains curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes. PID seeks to address two issues affecting biological processes presentation: (1) arbitrariness of pathway boundaries, (2) knowledge capture at different level of details. PID has adopted a network-level representation, similar to Reactome, HumanCyc, and KEGG. PID is focused on signaling and regulatory pathways and contains only structured data which it links to. The data in PID are from several sources: (1) highly curated "NCI-Nature Curated" collection of pathways, (2) Reactomes, (3) BioCarta.

PID: The Pathway Interaction Database.doc is a publication regarding the PID. The publication provides the description and requirements of the database. The provided information may be incomplete. You may wish to visit the PID website provided above for more information regarding PID. You may specify additional assumptions to address inconsistencies or missing information. These assumptions should be clearly specified in your submission.

References

1. Krupa, S., et al., *An Introduction to the NCI Pathway Interaction Database.* 2006.
2. Schaefer, C.F., et al., *PID: The Pathway Interaction Database*. 2008, Nature Precedings.

APPENDIX B: QUERIES (MINIMUM REQUIREMENT)

1. Count the number of *interactions* involved in "Syndecan-1-mediated signaling events" pathway (Pathway ID: 200036)

2. List the name, ID and type of all the molecules involved in "Syndecan-1-mediated signaling events" pathway (Pathway ID: 200036).

3. List the name and ID of all the interactions that "Syntenin" is involved in.

4. List the name, ID and modification of all the components in complex "Syndecan-1/HGF/MET"

5. Find the name, ID and role of all the molecules involved in modification interaction with interaction_ID 201725.