Project 2, CMPE 321, Introduction to Database Systems, Spring 2021

Due: **May 18, Tuesday, 23:55**

Ask the project related questions in the Moodle forum!

1 Introduction

The study of drugs or chemicals and the effects they have on a living organism is called pharmacology. A drug is a chemical that interacts with a target molecule (such as a protein), which in turn results in a change in the molecule's or the corresponding cell's behavior.

Drugs, in general, are not specific to only one particular molecule or cell. They may also interact with other molecules or other drugs and this is what causes side effects. For instance, anticancer drugs are selective for very rapidly dividing cells such as cancer cells and stimulate the cells to die. Hair loss is one of the side effects of anticancer drugs, since hair cells are also rapidly dividing. [1]

To cope with the tremendous number of drugs and targets (e.g., proteins that bind to drugs), online databases are created and commonly used in the literature. Online databases allow easy and fast access to information and most of them focus on a specific field. For instance, DrugBank contains information on drugs and drug targets, UniProt contains proteins and their structures, SIDER contains drugs and their side effects, whereas BindingDB curates the interactions between drugs and targets. Thus, a free online resource that integrates these databases would benefit drug discovery researchers, chemists, pharmacists, students, and the general public.

2 Project Description

In this project, you will design a unified Drug - Target database, called ***DTBank***. You will begin with a detailed description of the content. Then you will need to systematically go through parts of the standard database design process as you learned about in class, including conceptual design, logical design, and schema refinement.

2.1 *DTBank*: Structure

The *DTBank* should contain the following information:

1. **User** includes the following attributes; username, institute name, and password. There exists only one user with a specific username and institute. There can be an unlimited number of users.

2. **Database manager** consists of the following attributes: username and password. There exists only one database manager with a username. There can be at most 5 database managers registered to the system.

3. **DrugBank** includes DrugBank ID, drug name, description, and interaction with other drugs. By definition, each DrugBank ID is unique.

4. **SIDER** includes UMLS CUI (side effect IDs), DrugBank ID, and side effect name. By definition, each UMLS CUI is unique.

5. **BindingDB** includes Reaction ID, DrugBank ID, UniProt ID, target (protein) name, SMILES (chemical notation of drug), affinity in nM (the strength of the binding interaction between drugs and targets), the measure of the interaction (Ki, Kd, IC50), DOI (link on the web to identify the article or document that mentions the drug-target interaction. E.g., https://doi.org/10.1093/bioinformatics/bty593), authors of the article or document, the username of the first author, and institution of the first author. By definition, each Reaction ID is unique and the first author is a user of the *DTBank*.

[1] You can check out this link for further information on pharmacology.

6. **UniProt** includes UniProt ID and amino acid sequence of the corresponding protein. By definition, each UniProt ID is unique.

2.2 *DTBank*: Real data

Here, we provide a sample database according to previously mentioned entries, more data will be shared through Moodle. Table 1 shows 5 different users with their usernames, passwords, and institution names. Table 2 includes the usernames and passwords of database managers. Table 3 shows 7 different side effects of 2 distinct drugs. For instance, drugs with the ID of $DB00600$ and $DB00210$ might cause a burning sensation with the ID of $C0085624$. Table 4 shows a protein with the ID of $P10826$ and long the amino acid sequence of $MTTSGHACP...VSQSPLVQ$. Table 5 shows several entries from DrugBank. For instance, Monobenzone has an ID of $DB00600$, interacts with a drug whose ID is $DB15874$, and it is medically used for depigmentation. Finally, Table 6 shows entries from BindingDB. An example Drug - Target interaction is listed as follows: interaction with the ID of 51001683 occurred between the drug with the ID of $DB00600$ and the protein with the ID of $P15207$. The name of the protein is *Androgen receptor*. $DB00600$ drug has $Oc1ccc(OCc2ccccc2)cc1$ as the chemical notation. The IC50 value of binding affinity, that is the strength of the interaction between drug - target pair, is 234422.9 nM. The depicted interaction takes place in the article with a DOI of $10.1021/jm050403f$, one can easily check the article using this link doi.org/10.1021/jm050403f. Several authors contributed to this article such as M.A. Lill, F. Winiger, A. Vedani, and B. Ernst, from the University of Basel. Finally, the username of the first author is Lill, from the University of Basel.

username	**institution**	**password**
Charpentier	CIRD GALDERMA	charpentier99!.
Diaz	TBA	diaz.p25
Morgan	Amgen Inc	morgan.re.123
Lill	University of Basel	ma_lill000
Oyku.yilmaz	Bogazici University	oyku.yilmaz.10

Table 1: Sample data from User

username	**password**
selen.parlar	selen.parlar
riza.ozcelik	riza.ozcelik0
arzucan_ozgur	arzucan_135

Table 2: Sample data from Database Manager

umls_cui	**drugbank_id**	**side_effect_name**
C0085624	DB00600	Burning sensation
C1325847	DB00600	Sensitisation
C0152030	DB00600	Skin irritation
C0521491	DB00210	Application site pain
C0085624	DB00210	Burning sensation
C0009763	DB00210	Conjunctivitis
C0036572	DB00210	Convulsion
C0011603	DB00210	Dermatitis
C0152030	DB00210	Skin irritation

Table 3: Sample data from SIDER

uniprot_id	**sequence**
P15207	MEVQLGLGR...KPIYFHTQ
P10826	MTTSGHACP...VSQSPLVQ

Table 4: Sample data from UniProt

drugbank_id	name	drug_interactions	description
DB15874	Agalsidase alfa	['DB00608', 'DB01118', 'DB00600', 'DB00798']	Agalsidase alfa is a recombinant human α-galactosidase A similar to agalsidase beta. While patients generally do not experience a clinically significant difference in outcomes between the two drugs, some patients may experience greater benefit with agalsidase beta. Use of agalsidase beta has decreased in Europe, in favor of agalsidase alfa, after a contamination event in 2009. Agalsidase alfa was granted EMA approval on 3 August 2001.
DB00600	Monobenzone	['DB15874']	Monobenzone is the monobenzyl ether of hydroquinone used medically for depigmentation. Monobenzone occurs as a white, almost tasteless crystalline powder, soluble in alcohol and practically insoluble in water. It exerts a depigmenting effect on skin of mammals by increasing the excretion of melanin from the melanocytes. It may also cause destruction of melanocytes and permanent depigmentation.
DB00210	Adapalene	['DB00936', 'DB11085']	Acne vulgaris is a multifactorial disorder of the pilosebaceous unit involving increased sebum production, inflammation, and hyperproliferation/hyperkeratinization of the follicular infundibulum. It is also associated with Cutibacterium acnes. Adapalene is a third-generation topical retinoid used for the treatment of acne vulgaris. Adapalene has similar efficacy but a superior safety profile compared to tretinoin. [Tazarotene] is more efficacious than adapalene but is designated as pregnancy category X and hence is contraindicated in pregnant women. Adapalene can also be combined with benzoyl peroxide (BPO), which possesses bactericidal properties, and either adapalene alone, or adapalene BPO combination products, are commonly used to treat mild-to-severe acne.

Table 5: Sample data from DrugBank

reaction_id	drugbank_id	uniprot_id	target_name	smiles	measure	affinity_nM	doi	authors	username	institution
50078054	DB00210	P10826	Retinoid receptor	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	34.0	10.1021/jm00026a006	Charpentier, B; Bernardon, JM; Eustache, J; Millois, C; Martin, B; Michel, S; Shroot, B	Charpentier	CIRD GALDERMA
50078097	DB00210	P13631	Retinoid receptor	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	130.0	10.1021/jm00026a006	Charpentier, B; Bernardon, JM; Eustache, J; Millois, C; Martin, B; Michel, S; Shroot, B	Charpentier	CIRD GALDERMA
50078117	DB00210	P10276	Retinoid X receptor gamma/ retinoic acid receptor alpha	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	1100.0	10.1021/jm00026a006	Charpentier, B; Bernardon, JM; Eustache, J; Millois, C; Martin, B; Michel, S; Shroot, B	Charpentier	CIRD GALDERMA
50726488	DB00210	P10276	Retinoid X receptor gamma/ retinoic acid receptor alpha	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	1100.0	10.1016/S0960-894X(97)00405-8	Diaz, P; Michel, S; Stella, L; Charpentier, B	Diaz	TBA
50726496	DB00210	P13631	Retinoid receptor	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	130.0	10.1016/S0960-894X(97)00405-8	Diaz, P; Michel, S; Stella, L; Charpentier, B	Diaz	TBA
50726497	DB00210	P10826	Retinoid receptor	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	Ki	34.0	10.1016/S0960-894X(97)00405-8	Diaz, P; Michel, S; Stella, L; Charpentier, B	Diaz	TBA
50772412	DB00210	O15439	Multidrug resistance-associated protein 4	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	IC50	133000.0	10.1093/toxsci/kft176	Morgan, RE; van Staden, CJ; Chen, Y; Kalyanaraman, N; Kalanzi, J; Dunn, RT; Afshari, CA; Hamadeh, HK	Morgan	Amgen Inc
50773523	DB00210	O95342	Bile salt export pump	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	IC50	133000.0	10.1093/toxsci/kft176	Morgan, RE; van Staden, CJ; Chen, Y; Kalyanaraman, N; Kalanzi, J; Dunn, RT; Afshari, CA; Hamadeh, HK	Morgan	Amgen Inc
50775204	DB00210	O15438	Canalicular multispecific organic anion transporter 2	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	IC50	133000.0	10.1093/toxsci/kft176	Morgan, RE; van Staden, CJ; Chen, Y; Kalyanaraman, N; Kalanzi, J; Dunn, RT; Afshari, CA; Hamadeh, HK	Morgan	Amgen Inc
50777179	DB00210	Q92887	Canalicular multispecific organic anion transporter 1	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)-c1ccc2cc(ccc2c1)C(O)=O	IC50	133000.0	10.1093/toxsci/kft176	Morgan, RE; van Staden, CJ; Chen, Y; Kalyanaraman, N; Kalanzi, J; Dunn, RT; Afshari, CA; Hamadeh, HK	Morgan	Amgen Inc
51001683	DB00600	P15207	Androgen receptor	Oc1ccc(OCc2ccccc2)cc1	IC50	234422.9	10.1021/jm050403f	Lill, MA; Winiger, F; Vedani, A; Ernst, B	Lill	University of Basel

Table 6: Sample data from BindingDB

2.3 Part 1: Conceptual database design

Your task in Part 1 is to perform the Conceptual Database Design (or ER Design) – draw ER diagrams to capture all the information, following the approach described in lectures. While there are many ER-model variants, for this project, we expect you to use the ER notation from the **textbook and lecture**.

To receive full points for this part, you need to identify all the entity sets and relationship sets in a reasonable way. We expect there to be multiple correct solutions since the ER design is subjective. Your goal should be to reasonably capture the given information. For the entity set names, relationship set names, and attribute names that you will be using in your ER diagram, you can use the ones we have provided in Section 2.1 and Section 2.2. You can use underscores, spaces, numbers, uppercase, or lowercase letters to construct those names. It is required to use the features of ER modeling that you have learned from the lectures, including participation constraints, key constraints, weak entities, class hierarchy, and aggregation. You should get concrete data because it may help you understand the problem better. You should use online drawing tools, handcrafted diagrams will not be accepted.

2.4 Part 2: Logical database design

For the second part of the project, your task is to convert the ER diagrams into relational tables, based on the set of simple rules as described in the textbook and in lectures. You should provide the schema of each relation including the relation name, attribute names, and attribute domains.

2.5 Part 3: Schema refinement and normalization

For the third part of the project, your task is to analyze your design in Part 2 in terms of functional dependencies (FDs) and normal forms, then, refine your design if needed. You should explicitly list all of the non-trivial FDs. Then, for each relation you should determine if it is in **Boyce-Codd Normal Form** (BCNF) and you should explain how the requirements of BCNF are met (or not met) in terms of FDs. If a relation is not in BCNF, you should check whether it is 3NF and explain how the requirements for 3NF are met (or not met). If a relation is not in BCNF, you should either decompose it into BCNF relations or provide a justification if you decide not to decompose it. If you decompose a relation, you should explain whether the decomposition is lossless-join and dependency preserving. It is possible that your initial schema is already in BCNF. If this is the case, you still need to explain how the requirements of BCNF are met in terms of FDs for each one of your relations.

2.6 Part 4: Write SQL statements for the normalized schema

You are required to write SQL DDL statements that create the tables you designed for this part. You should specify all the constraints such PK, FK, Unique, NOT NULL, and other general constraints with CHECK. You should turn in two files:

1. createTables.sql
2. dropTables.sql

Make sure that you include a comment in each file.

3 Submission & Remarks

This project can be implemented either individually or as a team of two people. You are free to change teams in the upcoming projects. Place all `.sql` files and a PDF file contains the outputs of Part 1, Part 2, and Part 3 into a folder named with the student IDs of the team members separated by an underscore (e.g. `2017400200_2018700120`). Zip the folder for submission and name the `.zip` file with the same name. Submit the `.zip` file through Moodle until the deadline. **Any other submission method and late submissions are not allowed**.