**Übungen zu Wissenschaftliche Methodik I: Datenbanken (8.11.2023, 08:00-09:30)**
**Wintersemester 2023/24, M.Sc.-Studiengang Technische Biologie**

M.Sc. Jan Range
Institut für Biochemie und Technische Biochemie, Universität Stuttgart
E-Mail: jan.range@simtech.uni-stuttgart.de
Telefon: +49 711 685 60095

## Aufgabe 1: Datenbanksysteme und Transaktionen

Erläutern Sie mindestens zwei Vorteile von (relationalen) Datenbanken gegenüber Microsoft Excel-Dateien (oder Tabellenkalkulationsdateien im Allgemeinen).

## Aufgabe 2: Das relationale Datenmodell

Proteinsequenzen und Proteinstrukturen werden in öffentlichen Datenbanken über *Accessions* eindeutig identifiziert. Beispiel: Der Sequenz `Q9F4L3` aus der Uniprot-Datenbank wird die Struktur `2AG0` aus der PDB (Protein Data Bank) zugeordnet.

a. Gegeben sei eine Tabelle für Proteinstrukturen (`STRUCTURE`) mit den Attributen `NAME`, `PDB_ACCESSION` und `RESOLUTION` (Auflösung in Å). Welche Datentypen vergeben Sie für diese drei Attribute?

b. Erweitern Sie das Datenmodell um eine Tabelle für Proteinsequenzen (`SEQUENCE`) mit den Attributen `NAME` und `ACCESSION` (z.B. Uniprot-Accession). Wie können Sie den Sequenzen die Strukturen zuordnen und umgekehrt (1:1-, 1:n- oder n:n-Beziehung)?

c. Wie viele Tabellen benötigen Sie insgesamt in Ihrem Datenmodell für Protein-strukturen und Proteinsequenzen?

d. Ergänzen Sie Ihr Datenmodell entsprechend und erstellen Sie eine Skizze mit den notwendigen Primär- und Fremdschlüsseln (*primary keys* und *foreign keys*).

## Aufgabe 3: Abfragen mit SQL

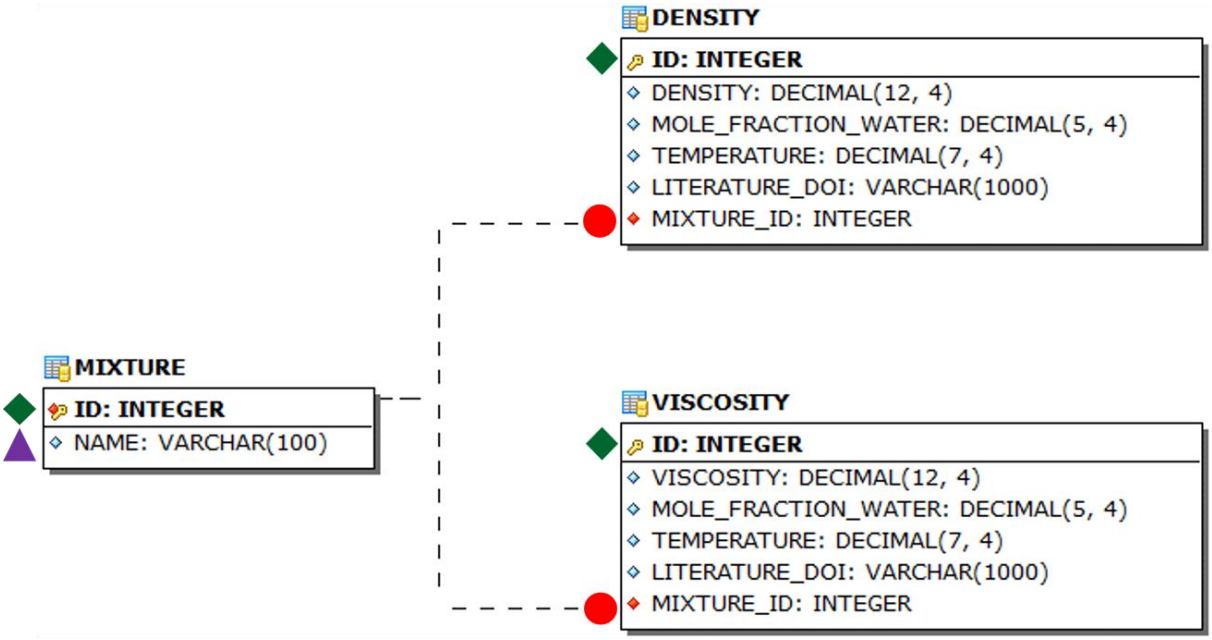


**Abbildung 1** Skizze zum Datenmodell der Datenbank *mixtures* mit drei Primärschlüsseln (*primary keys*, Rauten) namens `ID` und zwei Fremdschlüsseln (*foreign keys*, Kreise) namens `MIXTURE_ID`. Das Attribut `NAME` in der Tabelle `MIXTURE` darf nur eindeutige Werte haben (*unique*, Dreieck).

a. Formulieren Sie einen SELECT-Befehl in SQL, mit dem Sie in der Datenbank *mixtures* (Abb. 1) alle Viskositäten finden können, die bei Temperaturen von mindestens 25°C und einem Stoffmengenanteil für Wasser kleiner als 0,5 bestimmt wurden. Beachten Sie dabei, dass die Datenbank Kelvin als Temperatureinheit verwendet.

b. Ändern Sie Ihren vorherigen SELECT-Befehl, um folgende Frage zu beantworten: *Wie viele Publikationen* gibt es über Viskositäten, die bei mindestens 25°C und einem Stoffmengenanteil für Wasser kleiner als 0,5 bestimmt wurden?

c. Formulieren Sie einen SELECT-Befehl in SQL, mit dem Sie aus der Datenbank *mixtures* alle Dichten für Methanol-Wasser-Mischungen finden können. Schreiben Sie Ihren SELECT-Befehl so, dass mit dem Namen der Mischung (`'methanol-water'`) und nicht mit der `ID` der Mischung gesucht wird.

## Aufgabe 4: Datenbankanwendung in R

Für diese Aufgabe benötigen Sie die Bibliothek RODBC und Zugang zur Datenbank *mixtures*, am einfachsten über R-Studio auf https://davinci.ibvt.uni-stuttgart.de (erreichbar über VPN).

Beispiel-Befehle in R:

```
library(RODBC)
connection <- odbcConnect(dsn="mixtures", believeNRows=FALSE)
data <- sqlQuery(connection, "SELECT * FROM MIXTURE")
```

Die einzelnen Spalten sind in R über ihre Namen erreichbar, z.B.: `data$NAME`

a. Wählen Sie aus der der Tabelle `VISCOSITY` in der Datenbank *mixtures* alle Werte von Viskosität und Temperatur für die Mischung Ethylenglykol-Wasser (`'ethylene glycol-water'`), die bei einem Stoffmengenanteil für Wasser von 0.5 bestimmt wurden.

b. Zeigen Sie mit R einen Plot, der für diese ausgewählten Daten die Abhängigkeit der Viskosität von der Temperatur darstellt. Ergänzen Sie den Plot mit sinnvollen Achsenbeschriftungen.

c. Erstellen Sie einen zweiten Plot, diesmal mit logarithmierter Viskosität.

d. Bestimmen Sie für die Daten von beiden Plots jeweils die Korrelationskoeffizienten mit der Methode nach Pearson und Spearman.

e. Interpretieren und vergleichen Sie die Ergebnisse hinsichtlich der Annahmen für die Korrelationskoeffizienten nach Pearson und Spearman.