Opdracht: GenCluE

Deze opdracht bestaat uit twee delen en vormt samen met het tentamen de afronding van het vak Datamodellering en -verwerking 8C020. De opdracht toetst de stof die wordt behandeld in de hoofdstukken 1 tot en met 6 van het collegedictaat. De opdracht wordt gemaakt in groepen van twee personen. De uitwerkingen van **deel A** dienen op papier te worden ingeleverd uiterlijk op **donderdag, 1 november 2007, 18:00 uur**. Na beoordeling worden deze uitwerkingen teruggegeven op een van de volgende colleges. De definitieve uitwerkingen van deel A en de uitwerkingen van deel B bestaande uit een verslag en een Access database dienen uiterlijk **dinsdag, 20 november 2007, 18:00 uur** ingeleverd te worden. Het verslag wordt op papier ingeleverd en de database als Access database bestand op diskette of CD-ROM.

Deel A

De Gene Clustering en Expression (GenCluE) groep die bestaat uit onderzoekers van de universiteiten in Eindhoven en Maastricht onderzoekt genexpressie met behulp van een benadering die gebaseerd is op clustering van genen. Uw taak is om een computer applicatie te schrijven die gebaseerd is op een relationele database en die (gedeeltelijk) het werk van de groep automatiseert. De onderstaande informatie is bedoeld om u te helpen om de gebruikerseisen te definieren.

Aan elk gen correspondeert een sequentie van nucleotiden (een string die alleen uit de letters A, C, T en G bestaat). (Wij nemen aan dat de sequentie hoogstens 30 letters lang is). Als resultaat van de experimenten en met behulp van geautomatiseerde classificatie zijn de genen verdeeld in disjuncte clusters. Een van de belangrijkste onderzoeksvragen is om te checken of de genen uit dezelfde clusters ook verwante processen in de cell controleren. Daarvoor is aan elk gen een verzameling van sleutelwoorden die zijn functie beschrijven toegekend. Bijvoorbeeld, is het op die manier mogelijk om te checken of er woorden zijn die alleen in de beschrijvingen van genen uit een cluster voorkomen. Voor elk gen zijn er enkele URLs naar entries over het gen in verschillende databases.Verder zijn er voor elk gen mogelijke referenties naar wetenschappelijke publicaties waar men extra informatie over dat gen kan vinden. Er zijn er verschillende soorten publicaties (boeken, tijdschriften, proceedings, proefschriften, enz.). Een andere manier om biologisch betekenisvolle informatie uit de expressiepatronen te krijgen is het expressiegedrag van bekende familes van genen te onderzoeken. Er zijn 26 van zulke families. Van belang zijn ook eiewitten die zijn gecodeerd door nucleotidesequenties die met de genen zin gassocieerd. Voor deze eiwitten is het handig om, behalve de basis gegevens (b.v., naam, aminozuur sequentie), links (URLs) te hebben naar publieke databases waar men nadere informatie (b.v. 3D structuren van eiwitten) kan vinden.

Onderstaande vragen leiden tot een Entity Relationship Model, een relationeel database model en een aantal SQL queries.

1. De gegeven omschrijving van GenCluE is niet helemaal volledig en op sommige punten onduidelijk. Geef in natuurlijke taal een duidelijke omschrijving van de informatiebehoefte van de gebruiker. Ontbrekende informatie mag je zelf aanvullen.
2. Geef een voorbeeld van referentie (wetenschappelijke publicatie) met bijbehorende velden (attributen) die u in de GenCluE database op zou kunnen slaan.
3. Maak een Entity-Relationship diagram voor de genen en bijbehorende attributen. Geef ook een toelichting op de gemaakte keuzes.
4. Maak nu op basis van de informatiebehoefte (vraag 1) het Entity-Relationship Model voor GenCluE. Het ER-diagram van vraag 3 is hiervan een onderdeel. Geef ook de primaire sleutels en de cardinaliteiten van de relaties. Geef behalve het ER-diagram ook een toelichting.

De database moet ook gegevens over de leden van de groep bevatten. De leden zijn verdeeld in onderzoekers en ondersteunend personeel. Daarnaast willen wij ook records over elke experiment bijhouden. Deze records bevatten infromatie over de tijd waarop het experiment uitgevoerd is, de betrokken genen en de deelnemende groepsleden.

5. Breid het ERM zodanig uit dat de nieuwe informatiebehoeftes in het model zijn inbegrepen.
6. Zijn er attributen waarvoor in sommige gevallen meer dan één waarde wordt bijgehouden? Pas het Entity-Relationship Model aan zodat er geen meerwaardige attributen voorkomen.
7. Specificeer van drie verschillende attributen met verschillende domeinen geschikte domeinconstraints.
8. Maak op basis van het ERM uit vraag 6 een relationeel databasemodel.

Gegeven het relationeel database model uit vraag 8 is het mogelijk om queries te formuleren waarmee je relevante informatie uit de database kunt opvragen.

9. Voor een successvol onderzoek is het belangrijk om een overzicht van het gerelateerde werk te hebben. Schrijf een SQL query die een lijst geeft van alle artikels die geschreeven zijn tussen 2002 en 2007 en die over genen van een famille gaan waarvan de beschrijving een bepaald (sleutel)woord *w* bevat (b.v. athelrosclerosis).
10. Vaak hebben genen die verwante sequenties hebben ook verwante functionaliteit. Zoals al gezegd, kan men conclusies trekken gebaseerd op de beschrijving (sleutelwoorden) van de genen. Bedenk een SQL query die voor een gegeven nucleotidensequentie *s* (b.v. "ATTAGTGCC") alle sleutelwoorden geeft die voorkomen in de beschrijving van de genen die de sequentie *s* bevatten. Schrijf daarna een query die de URLs geeft van alle eiwitten die gecodeerd zijn door genen die de sequentie *s* bevatten.
11. Gegeven een gen cluster *c* (e.g. Collagen) en een groeplid *m* (b.v. Jansen) schrijf een SQL query die alle experimenten die in de laatste vijf maanden zijn uitgevoerd geeft zodanig dat *m niet* betrokken is geweest in die experimenten en de experimenten te maken hebben met de genen van cluster *c*.

Deel B

Op basis van het relationeel database model en de mogelijke queries uit deel A kunnen we nu een realisatie maken van GenCluE met behulp van Microsoft Access. We zullen hiertoe gebruik maken van *tabellen*, *queries*, *formulieren*, *rapporten* en een *switchboard.*

12. Ontwerp in Access alle tabellen uit het relationeel database model (vraag 8) en geef de relaties tussen de tabellen aan.
13. Ontwerp een formulier om gegevens over genen in te kunnen voeren.
14. Ontwerp een formulier voor het toevoegen van groepsleden aan de database.
15. Ontwerp een formulier voor het toevoegen van weteshcappelijke publicaties aan de database.
16. Ontwerp queries (m.b.v. QBE) die equivalent zijn (geven dezelfde resultaten als) aan die van vraag 9.
17. Ontwerp queries (m.b.v. QBE) die equivalent zijn aan de queries van vraag 10.
18. Ontwerp een rapport om de groepsleden die onderzoeker zijn netjes af te drukken.
19. Ontwerp een query die equivalent is aan die van vraag 11 en een bijbehorende rapport.
20. Maak met behulp van een *switchboard* een origineel startscherm voor uw applicatie.
21. Ontwerp meer formulieren, queries, en rapporten die nuttig kunnen zijn voor GenCluE.
22. Vind op de web enkele databases van genen (b.v. GenBank, KEGG) en schrijf een korte vergelijking met uw database.

Uw veslag moet ten minste de namen van de queries, reports en forms, en een korte disussie voor elk uitwerking van de bovengenoemde vragen bevatten. Vul a.u.b. de databases met gegevens (niet te veel!) zodanig dat de implementatie (queries, forms en reports) getest kan worden.

Succes!