EXAMEN: **Scriptingtalen**

Prof. Dr. Peter Dawyndt
1e Bachelor Informatica
groep 1

woensdag 22-08-2007, 8:30h
academiejaar 2006-2007
tweede zittijd

Opgave 1

Gevraagd wordt een **awk** script `procentGC.awk` te schrijven dat voor elke *DNA sequentie* uit een gegeven bestand of reeks bestanden in *FASTA formaat* het *%GC* berekent en uitschrijft naar standaard uitvoer. Finaal moet ook het minimum, maximum en gemiddeld %GC van alle sequenties worden uitgeschreven. Zorg ervoor dat het script een functie `berekenGC` bevat die voor een gegeven DNA sequentie, die als argument aan de functie wordt doorgegeven, het %GC berekent. Roep waar nodig deze functie aan in het script. Onderstaande sessie toont de uitvoer die het script moet genereren wanneer het wordt gebruikt om de opgegeven bestanden `seq1.fasta` en `seq2.fasta` te verwerken. Hierbij bevat de eerste kolom de speciesnaam (het derde veld uit de beschrijvingsregel), waarvoor 30 lettertekens worden gereserveerd. De tweede kolom geeft het %GC weer, afgerond tot op twee cijfers na de komma. De derde kolom bevat het accession number (het tweede veld uit de beschrijvingsregel). Termen die hierboven met een cursief lettertype werden weergeven, worden hierna in detail uitgelegd.

```
$ awk -f procentGC.awk seq1.fasta seq2.fasta
Bacillus cereus                53.49% (DQ207729)
Burkholderia xenovorans        56.41% (U86373)
Clostridium acetobutylicum     52.68% (U16165)
Geobacter metallireducens      56.40% (L07834)
Listeria welshimeri            53.61% (X98532)
Methanosarcina acetivorans     56.63% (M59137)
Oceanobacillus iheyensis       52.85% (AB010863)
Thermus thermophilus           63.96% (X07998)
Xanthomonas campestris         55.13% (X95917)
Bacillus sporothermodurans     54.38% (U49078)

minimum                        52.68%
maximum                        63.96%
gemiddelde                     55.55%
```

DNA sequentie: Een DNA sequentie kan worden voorgesteld door een tekenreeks die enkel bestaat uit de lettertekens A, G, C en T (maar in de praktijk worden ook nog andere letters gebruikt, die onzekere baseparen of gaten aangeven in de sequentie). Het %GC wordt berekend als de verhouding van het aantal lettertekens `G` en `C` in de tekenreeks, ten opzichte van het totaal aantal `A`'s, `G`'s, `C`'s en `T`'s (andere lettertekens worden genegeerd). Voor `GCCTGCAG` is het %GC = 75%.

%GC: In de genetica is de guanine-cytosine inhoud (%GC) een karakteristieke eigenschap van het genoom van een gegeven organisme of van elk ander stuk DNA of RNA. Normaalgezien wordt deze eigenschap uitgedrukt als een percentage, en geeft ze de verhouding weer van de GC baseparen in de DNA molecule of genoomsequentie die onderzocht wordt. G staat hierbij voor guanine en C voor cytosine. De overblijvende baseparen van een DNA molecule zullen dan bestaan uit de basen A (adenine) en T (thymine), zodat de berekening van het %GC op een indirecte manier ook de berekening van het %AT oplevert (%GC = 58% $\Rightarrow$ %AT = 42%). GC-baseparen zijn in het DNA verbonden met drie waterstofbindingen in plaats van twee in het geval van de AT-baseparen. Dit zorgt ervoor dat GC-paren sterker en beter resistent zijn tegen denaturatie bij hoge temperaturen, waardoor het %GC dus meestal groter is bij hyperthermofielen.

FASTA formaat: FASTA is een tekstgebaseerd bestandsformaat dat gebruikt wordt in de bioinformatica om DNA of eiwitsequenties op te slaan. Individuele baseparen of eiwitresidu's worden daarbij voorgesteld door één-letter codes. Het formaat laat ook toe om de verschillende sequenties te laten voorafgaan door sequentienamen en andere informatievelden.
Een sequentie in FASTA formaat begint met een één-regel beschrijving, gevolgd door de eigenlijke sequentiegegevens die eventueel kunnen gesplist worden over verschillende regels. De regel met de beschrijving wordt onderscheiden van de sequentiegegevens door een "groter dan" symbool (">") in de eerste kolom. Het woord dat volgt op het ">" symbool is de identifier van de sequentie, en de rest van de regel is de beschrijving (beide zijn optioneel). Er mag geen spatie staan tussen het ">" symbool en de eerste letter van de identifier en het is aanbevolen dat alle regels korter zijn dan 80 lettertekens. Elke sequentie eindigt waar een nieuwe regel begint met een ">" symbool, wat de start van een nieuwe sequentie aangeeft, of op het einde van het bestand. Een eenvoudig voorbeeld van één enkele sequentie in FASTA formaat:

```
>118480563|DQ207729|Bacillus cereus|16S ribosomal RNA gene, complete sequence
AGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAATGGATTA
AGAGCTTGCTCTTATGAAGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCCATAAGACTGGG
ATAACTCCGGGAAACCGGGGCTAATACCGGATAACATTTTGAACCGCATGGTTCGAAATTGAAAGGCGGC
TTCGGCTGTCACTTATGGATGGACCCGCGTCGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCAA
CGATGCGTA
```

De beschrijvingsregel — de regel die begint met een ">" symbool — geeft een naam en/of een unieke identifier aan de sequentie, en bevat vaak ook nog andere informatievelden. Informatievelden worden dan van elkaar gescheiden door het "|" symbool. Verschillende sequentiedatabanken maken gebruik van gestandaardiseerde beschrijvingsregels, wat helpt om automatisch de informatie uit de regels te kunnen extraheren. Een voorbeeld van een FASTA bestand met meerdere sequenties, gegenereerd voor de GenBank databank van het NCBI, ziet er als volgt uit:

```
>571435|U16165|Clostridium acetobutylicum|NCIMB8052 16S ribosomal RNA (rrn) gene
TGGCGGCGTGCTTAACACATGCAAGTCGAGCGATGAAGCTCCTTCGGGAGTGGATTAGCGGCGGACGGGT
GAGTAACACGTGGGTAACCTGCCTCATAGAGGGGAATAGCCTTTCGAAAGGAAGATTAATACCGCATAAG
ATTGTAGTGCCGCATGGCATAGCAATTAAAGGAGTAATCCGCTATGAGATGGACCCGCGTCGCATTAGCT
AGTTGGTGAGGTAACGGCTCACCAAGGCGACGATGCGTAGCCGACCTGAGAGGGTGATCGGCCACATTGG
GACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTG
>996091|L07834|Geobacter metallireducens|16S ribosomal RNA gene
AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGAGTGCCTAACACATGCAAGTCGAACGTGAAGGGGG
CTTCGGTCCCCGGAAAGTGGCGCACGGGTGAGTAACGCGTGGATAATCTGCCCAGTGATCTGGGATAACA
TCTCGAAAGGGGTGCTAATACCGGATAAGCCCACGGAGTCCTTGGATTCTGCGGGAAAAGGGGGGGACCT
TCGGGCCTTTTGTCACTGGATGAGTCCGCGTACCATTAGCTAGTTGGTGGGGTAATGGCCCACCAAGGCT
ACGATGGTTAG
```

Na elke beschrijvingsregel volgen één of meerdere regels die de sequentie beschrijven. Deze regels moeten korter zijn dan 80 lettertekens. Sequenties kunnen zowel DNA-sequenties als eiwitsequenties voorstellen, en ze kunnen gaten bevatten die worden voorgesteld door een min-teken ("−").

Opgave 2

Implementeer een shell script dat een gebruikersnummer (uid) meekrijgt als argument, en dat als resultaat op afzonderlijke regels de naam, de home directory, de login shell en het groepsnummer van de corresponderende gebruiker wegschrijft naar standaard uitvoer. Vermeld de betekenis van de karakteristiek die op elke regel wordt weergegeven. Schrijf ook de naam van de groep die correspondeert met het groepsnummer weg naar standaard uitvoer, eventueel gevolgd door de andere groepen waartoe de gebruiker zou kunnen behoren (zie uitvoer van het commando `groups`).
Zorg ervoor dat het script altijd wordt uitgevoerd door de **bash** shell. Maak gebruik van de commando's `sed` en `cut` (niet van `awk` en `groups`) en van de bestanden `/etc/passwd` en `/etc/group`. Schrijf gepaste mededelingen naar standaard uitvoer indien niet juist één argument wordt meegegeven aan het shell script, indien het argument geen gebruikersnummer voorstelt, en indien het paswoordbestand of groepsbestand niet kan gelezen worden door de gebruiker van het script.

Opgave 3

Voor de DVD-verhuurketen die gebruikt maakt van de Sakila databank wordt gevraagd om SQL zoekopdrachten te formuleren die een antwoord bieden op de volgende vragen. Zorg er telkens voor dat de kolommen van de resulterende tabel een zinvolle naam krijgen. Geef op je antwoordblad ook aan hoeveel records de resultatentabel bevat.

1. Geef de titel, de lengte en de naam van de taal en de originele taal van elke film waarvan de gebruikte taal niet gelijk is aan de originele taal en die langer duurt dan 150 minuten. Zorg ervoor dat het resultaat gesorteerd wordt eerst in dalende volgorde van de lengte van de film, en daarna alfabetisch op filmtitel.

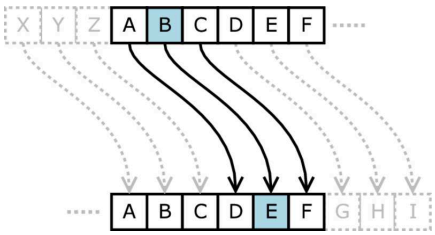
2. Geef de voor- en familienaam van alle klanten uit het klantenbestand van de DVD-verhuurketen die in Spanje woonachtig zijn.

3. Geef de titel van de films waarvan elke winkel uit de DVD-verhuurketen minstens 4 exemplaren in voorraad heeft.

Hint: De Sakila databank werd ingeladen in een MySQL RDBMS en is nog steeds bevraagbaar via de phpMyAdmin webapplicatie (`http://microlab.ugent.be/phpMyAdmin/index.php`). De guest-account die werd aangemaakt voor deze databank heeft als loginnaam *guest* en als paswoord *guest*.

Opgave 4

Het Caesarcijfer is een klassieke manier om tekstberichten te coderen (versleutelen) en te decoderen (ontsleutelen). Het is vernoemd naar Julius Caesar, die het gebruikte om te communiceren met zijn veldheren. De versleuteling werkt door elke letter van het alfabet te vervangen door een letter die enkele plaatsen verder in het alfabet voorkomt. Hierbij wordt het alfabet als circulair beschouwd, wat betekent dat na de letter `Z` opnieuw de letter `A` volgt. Vandaar dat ook de term rotatie of verschuiving gebruikt wordt bij deze operatie. Bij Rot3 (een rotatie van drie) wordt de letter A dus vervangen door de letter D (zie figuur).



Bij manuele codering en decodering van het Caesarcijfer is het gebruikelijk om bij de versleuteling en de ontsleuteling gebruik te maken van twee alfabetten. Een regulier alfabet dat gebruikt wordt in de originele tekst en een geroteerd alfabet dat gebruikt wordt in de gecodeerde tekst. In het geval van Rot3 zouden deze twee alfabetten er dus als volgt uitzien:

```
origineel: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
gecodeerd: D E F G H I J K L M N O P Q R S T U V W X Y Z A B C
```

Hierna vervangt men de letters van de originele tekst door de letter die eronder staat in de tabel. Zodoende wordt de geheime boodschap:

```
originele tekst:  D I T I S Z E E R G E H E I M
gecodeerde tekst: G L W L V C H H U J H K H L P
```

Het ontsleutelen van een gecodeerde tekst werkt dan precies andersom.

1. Declareer twee globale arrayvariabelen `Origineel` en `Gecodeerd`, waarin de 26 lettertekens van een regulier (resp. geroteerd) alfabet kunnen worden opgeslagen. Deze variabelen mogen niet toegankelijk zijn buiten de module waarin ze gedeclareerd worden.

2. Schrijf een procedure `Rotatie` die de variabele `Origineel` vult met het reguliere alfabet en de variabele `Gecodeerd` met een alfabet na rotatie door een geheel getal, dat als argument wordt doorgegeven aan de procedure. Deze procedure mag niet toegankelijk zijn buiten de module waarin ze gedeclareerd wordt.

3. Gebruik de procedure `Rotatie` voor het schrijven van een functie `Codeer` die een gegeven tekenreeks omzet naar de corresponderende Rot$n$-gecodeerde tekenreeks. Het getal n en de gegeven tekenreeks moeten als argument worden doorgegeven aan deze functie. Je mag hierbij uitgaan van het feit dat de tekst alleen hoofdletters bevat. Niet-alfabetische lettertekens moeten zonder wijziging worden overgenomen in de gecodeerde tekst.

4. Schrijf een duale functie `Decodeer`, die een Rot$n$-gecodeerde tekenreeks terug omzet in de originele tekenreeks.

5. Gebruik de voorgaande functies om in de derde kolom van werkblad `opgave4` Rot$n$-codering (waarbij n wordt gegeven in de tweede kolom) toe te passen op de tekst uit de eerste kolom. Pas daarna in de vierde kolom Rot$n$-decodering toe op de derde kolom, en controleer of het resultaat na decodering hetzelfde is als in de eerste kolom. Het eindresultaat moet er dan uitzien zoals in de onderstaande figuur.

originele tekst	rotatie	gecodeerde tekst	gedecodeerde tekst
DIT IS ZEER GEHEIM	3	GLW LV CHHU JHKHLP	DIT IS ZEER GEHEIM
VENI VIDI VICI	17	MVEZ MZUZ MZTZ	VENI VIDI VICI
DE OMNIS BELGAE FORTISSIMI SUNT	9	MN XVWRB KNUPJN OXACRBBRVR BDWC	DE OMNIS BELGAE FORTISSIMI SUNT
IACTA ALEA EST	13	VNPGN NYRN RFG	IACTA ALEA EST
GALIA EST PACTATA	22	CWHEW AOP LWYPWPW	GALIA EST PACTATA