Data Mining : les données de base

Stocker, manipuler, récupérer, formater des données ? perl, psql

Vous avez appris à créer et manipuler des données pour les stocker et éventuellement leur poser des questions. C'est ce que nous allons vérifier dans un premier temps.

Cas 1. Etude Marketting.

Vous êtes une société de production artistique. Vous désirez faire une enquête de satisfaction sur l'ensemble des spectacles de théâtre que vous avez produit cette année. Votre base de données relationnelle qui s'appellera MARKET doit intégrer le plus de données possibles sur les spectacles (artistes, lieu de représentation …) et sur les spectateurs. Bien sûr, vous devez prévoir de stocker l'appréciation donnée par les spectateurs qui auront le loisir de répondre soit à la sortie du spectacle dans une urne, soit en remplissant un formulaire sur le site web de votre société de production. Vous optez pour un logiciel de gestion de bases de données relationnelles.

a) Etablissez un schéma entités-relations qui répondent le mieux possible à ces attentes et qui vous permettra de créer la base de données effectivement.
b) Fournissez le schéma relationnel associé.
c) Quel logiciel utiliser ?
d) Et pour la création du formulaire en ligne sur votre site web ?

Nous supposons que votre base de données MARKET est réalisée. Une requête bien formulée vous donne la vue suivante appelée RESA : cette vue est disponible sous la forme de fichiers ASCII *resa.psql* pour la structure *resa.txt* pour les données sur le site http://www.math-info.univ-paris5.fr/sip-lab/lomn/Cours/DM/Material/ .

1. Transférer cette table dans votre base psql personnelle (commande psql puis \copy).
2. Quelle est la proportion de femmes et d'hommes qui ont réservé pour ce spectacle ?
3. A partir de votre base personnelle, utilisez l'environnement psql pour :
    1. en fait, il y a eu un problème de saisie. Les gens qui sont marqués 'o' pour le champ « Payé » n'ont pas payé et devrait être marqué 'n'. Les autres ont payé et devraient être marqués 'o'. Effectuez le changement.
    2. Quelle est la proportion de femmes parmi celles qui ont payé ayant apprécié le spectacle ? Quel est le jour ayant acueilli le plus de personnes ? le plus de réservations différentes ? le plus d'accueil favorable ?

Cas 2. Recherche génétique

Vous êtes un laboratoire de recherche en biologie moléculaire ou en génétique. Vous devez travailler sur les différents génomes séquencés à l'heure actuelle. Pour cela, vous désirez créer une base de données locale à votre laboratoire, appelée GENES, qui récupère des données réparties sur différentes bases de données publiques et que vous ferez évoluer en fonction de problématique propre de recherche. En utilisant la description succincte des données manipulées par un généticien fournie en Annexe A. ,

a) Etablissez un schéma entités-relations qui répondent le mieux possible à ces attentes et qui vous permettra de créer la base de données effectivement.
b) Fournissez le schéma relationnel associé.
c) Comment récupérer des données distantes ?

Nous supposons que votre base de données GENES est réalisée. Une requête bien formulée vous donne la vue suivante appelée GENE : cette vue est disponible par phpMyAdmin sur le site https://www.ens.math-info.univ-paris5.fr/phpPgAdmin . L'environnement de base de données *lomenie* est accessible sur le serveur opale pour tous.

1. Donnez les séquences de longueur supérieure à 600 bp (paire de bases).
(syntaxe SQL spécifique suivante par exemple : SELECT "length" FROM "GENES"."GENE" WHERE "length" > '1000')
2. Donnez les séquences ADN correspondantes.
3. Donnez les séquences comportant la suite 'TTTT'.
4. Donnez les séquences comportant la suite 'TT' puis n'importe **quel nucléotide** puis 'TT'.
5. Donnez les séquences comportant la suite 'TT' puis n'importe **quelle suite de nucléotides** puis 'TT'.
6. Récupérez sur le site ftp à l'aide d'une ligne de commande le fichier suivant *splice .tar.gz* que l'on peut trouver à l'URL suivante : ftp.cs.toronto.edu dans le répertoire pub/neuron/delve
7. Ecrivez le programme PERL suivant – « firstsearch.pl » - à l'aide de gedit puis exécutez-le : *perl firstsearch.pl*. Ensuite à l'aide de la commande **man**, étudiez la documentation pour PERL. Eventuellement, surfez sur le WEB.

```
#!/usr/bin/perl -w
# Look for nucleotide string in sequence data

my $target = "ACCCTG";
my $search_string= 'CCAAATTCTTCGGGACCCTGGGGGGGTTAAATTACCCTGACCCTGATG'.
    'CATGGTATGTACAGTAGACTAGGACAACCCTGGGGTAGA';

my @matches;

foreach my $i (0..length $search_string) {
    if ($target eq substr ( $search_string, $i, length $target)) {
      push @matches, $i;
    }
}
print "My matches occured at the following offsets : @matches.\n";
print "done\n";
```

Que faire avec des données ?

Cas 1. Etude Marketting

Nous supposons que votre base de données MARKET est réalisée. Une requête bien formulée vous donne la vue suivante appelée RESA :

Date	Nom	Nb	Payé	Commentaire	App.	Sexe.
16/06/2003	SILVERT Loic	1			A	M
16/06/2003	DE LA BARRE Juliette	3			A	F
16/06/2003	SILVERT Catherine	2			A	F
16/06/2003	FERNANDES Nathalie	4			B	F
16/06/2003	LEPRINCE Olivia	3			A	F
16/06/2003	TASCON Eve-Laure	1	o		B	F
16/06/2003	MIOSSEC Laurent	4			C	M
16/06/2003	LESOT Didier	7			A	M
16/06/2003	MESSAGER cecile	2			B	F
16/06/2003	GLORY Estelle	2			A	F
16/06/2003	PERON Guillaume	1			D	M
16/06/2003	STRAGIER François	1			C	M
16/06/2003	LACHAUD Béatrice	1			A	F
16/06/2003	ROSSIGNOL Raphaël	1			A	M
17/06/2003	RICHARD Fred	2			B	M
17/06/2003	GUILLOMAUD Armelle	1			C	F
17/06/2003	CLOPPET Florence	2			C	F
17/06/2003	ATALAY Volkan	1	o		B	M
17/06/2003	NARDY Nicole	1			C	F
17/06/2003	POULAIN Pascale	5		une place déjà payée	B	F
17/06/2003	NORMAND Jérôme	3			B	M
17/06/2003	BEVERINI Agnès	3			B	F
17/06/2003	MIDELET Annie	2		une place - de 25 ans	C	F
17/06/2003	PAUMIER Benoit	1			C	M
17/06/2003	MULLER Scott	1			C	M
17/06/2003	TOMESCU Anca	1			B	F
17/06/2003	MIGNOT Antoine	1			B	M
17/06/2003	ZAOUI Aissa	1	o		A	M
18/06/2003	Truccolo José&Marti	2			A	M
18/06/2003	TRUCCOLO Johann	1			A	M
18/06/2003	LAMORTHE Valerie	3			A	F
18/06/2003	DE LANVERSIN Anne	1			A	F
18/06/2003	DELPEUT Reine	3			B	F
18/06/2003	COLLOBERT M-Aline	2			A	F
18/06/2003	TISON Denis	2			B	M
18/06/2003	CHENE Marie	2			C	F
18/06/2003	BAUGHAN Michael	1			A	M
18/06/2003	ROCCA Claude	3		Payé une place	B	F
18/06/2003	MILAN JB	3			A	M
18/06/2003	PORTEVIN Liliane	3			B	F
18/06/2003	DE POMMERY Catherine	2			A	F
18/06/2003	VISIERE Juliette	2			D	F
18/06/2003	HORRI Selma	2		dont 1 réduit	B	F
18/06/2003	BONNET Dorothée et J	2			B	F
18/06/2003	BALLOY Linda	2			C	F
18/06/2003	CHESNOY Margareth	3			A	F
18/06/2003	DE LANVERSIN anne	1			A	F
18/06/2003	DEVILLERS André	2			A	M
18/06/2003	EMMA & MANU	2			C	M
18/06/2003	REFFAY Jean-Yves	2			A	M
18/06/2003	PLOUVIER Laurence	1		TR	A	F
18/06/2003	URBAIN Estelle	1		TR	B	F
19/06/2003	LY-YUNG Sandra	2			A	F
19/06/2003	HASKI Michel	2			A	M

1. Pouvez-vous faire parler ces données ? Sinon faites la liste de ce qui vous en empêche ? Si oui,
2. Quelles sont vos instruments mathématiques ou informatiques ?
3. Si vous en trouvez, modélisez le problème et déterminer les différentes étapes du processus.

Cas 2. Recherche génétique

Nous supposons que votre base de données GENES est réalisée. Une requête bien formulée vous donne la vue suivante appelée FAMILLE :

Id	Type	Id_gene	Long.	Seq
BX612931	cDNA	33502818	664	TGGTGACTGGTGGTCTCTATCA TACCTTTTGGTTGTGTTTTAGCTT GTAGGAAGAACGGG …
BX068059	mRNA	27641340	1002	TATTATAACAAACACGCCACCT TATTGAGTTTTTATTGATTCTAG AAGGTAAATATTTCG …
AJ284253	mRNA	6932132	429	GGAAGCCAGGCTGTCTCTCCCT CAAATAAAACCTAGCGG TTTACGGAATCTTCA …
BY612931	cDNA	33502818	523	TGGTGACTGGTGGTCTCTATCA TACCTTTTTTTATATTTCGCTT GTAGGAAGAACGGG …
BZ068059	mRNA	27641340	253	CAAACACGCCACCT T CAAATGTATCAAAACCTAGCGG GTATC …
CY612931	cDNA	33502818	1398	TCTCCACCT T CAAATGTATCATACCTTTT AACGGGTGGTGACTGGTGG …
DZ068059	cRNA	27641340	134	CAAACACTTTATATTTCGCTT GTAGGAAGACCTAGCGG GTATC …
AJ257253	cDNA	6932132	1523	GGAAGCCAGGCTGTCTCTCCCT TTTACGGAATCTTCA ATTGAGTTGGTTGTGT TTTATTGATTCTAG AAGGTAAAG …

1. Pouvez-vous faire parler ces données, et particulièrement la dernière colonne ? Sinon faites la liste de ce qui vous en empêche ? Si oui,
2. Quelles sont vos instruments mathématiques ou informatiques ?
3. Si vous en trouvez, modélisez le problème et déterminer les différentes étapes du processus.

Visualiser et formater des données ? gnuplot, openoffice, awk

Cas 1. Tableau de mesures 1D.

1. Voici un tableau de mesures purement numériques. Pouvez-vous prédire la valeur inconnue dans le couple (50, ?) ? Sinon faites la liste de ce qui vous en empêche ? Si oui,
    a) Quelles sont vos instruments mathématiques ou informatiques ?
    b) Si vous en trouvez, modélisez le problème et déterminer les différentes étapes du processus.

16	0,00
17	0,01677313
18	0,12393761
19	0,91578194
20	2,48935342
21	6,76676416
22	18,3939721
23	50
24	50
27	50
28	18,3939721
29	6,76676416
30	2,48935343
31	0,91578201
32	0,12393812
33	0,01677688
34	0,00229771
35	5,12E-04

40	0,22459823
41	0,6105213
42	1,65956895
43	4,51117611
44	4,51117611
45	12,262648
50	?
53	33,3333333
54	33,3333333
55	12,262648
56	4,51117611
57	4,51117611
58	1,65956895
59	0,6105213
60	0,22459823
63	0,00411366
64	0,00151333

2. Récupérer le tableau de mesures *1D.txt* décrivant une courbe sur le site web initial.
3. Quels sont les logiciels dont vous disposez pour visualiser cette courbe ?
4. Et visualisez la !

Cas 4. Tableau de mesures 2D.

1. Pouvez-vous indiquer des tendances observables dans les données 2D ci-dessous? Sinon faites la liste de ce qui vous en empêche ? Si oui,
    a) Quelles sont vos instruments mathématiques ou informatiques ?
    b) Si vous en trouvez, modélisez le problème et déterminer les différentes étapes du processus.

4,918	-0,256
5,005	-0,911
4,947	6,626
5,086	1,361
5,011	-8,983
5,053	-9,622
4,95	-4,036
5,075	0,631

5,084	0,309
5,062	-6,232
5,066	-9,566
4,991	-4,326
5,014	-5,706
4,916	-5,93
5,053	1,513
5,07	6,005

5,04	1,805	14,225	-2,45
4,963	-3,017	-0,81	1,883
5,049	6,684	11,362	-1,211
4,903	-0,828	-1,997	1,324
4,963	-2,023	11,061	2,975
4,939	3,933	1,953	0,335
4,987	6,892		

2. De même que précédemment, récupérer le tableau de mesures *2D.txt* décrivant un nuage de points 2D.
3. Le nuage n'est pas bien formaté. Il faut le reformater. Par exemple, certaines lignes comportent une troisième coordonnée égale à 0. Il faut l'enlever. On va utiliser la commande awk d'Unix.
   - Créez un programme *reformate.awk* à l'aide de gedit qui contient les deux lignes de code suivante :

     ```
     {print $1,$2}
     END{print NR}
     ```

   - Appliquez-le au fichier 2D.txt en lançant : *awk –f reformate.awk 2D.txt*
   - Créez un fichier reformaté avec cette formulation *: awk '{print $1,$2}' 2D.txt >2Df.txt*
4. Quels sont les logiciels dont vous disposez pour visualiser ce nuage ?
5. Et visualisez le !
6. Un nuage de points 2D discret peut aussi être vu comme une image 2D.
   1. Récupérez le programme écrit en langage C pour écrire des images au format pgm ou ppm (format ASCII) dans le fichier *pgm.tar.gz* sur le site précédent.
   2. Compilez ce programme pour construire l'exécutable.
   3. Etudiez le format des images fournies (commande more) et visualisez-les (commande *xv* ou *gimp*).
   4. Transformez votre nuage de points *2D.txt* en une image *2D.pgm (points noirs sur fond blanc)* .

Annexe A.

Depuis l'avènement du Human Genome Project, l'humanité peut consulter son patrimoine génétique : les séquences de nucléotides de chacun de ses 23 chromosomes et leur décomposition en gènes (http://www.ncbi.nlm.nih.gov/genome/guide/human/).

Les données afférentes à ces recherches sont innombrables et de natures hétérogènes. L'ensemble des bases de données également.

Si l'on considère une base de données dédiées aux séquences nucléotides spécifiquement comme GenBank par exemple, il faut considérer qu'une séquence appartient à une espèce, qu'elle peut contenir plusieurs gènes, qu'elle provient d'un tissu particulier. Par ailleurs, sa découverte a été confirmée par des publications dans des revues scientifiques. On donne un exemple de fichier de description obtenu pour une séquence extraite d'une de ces bases de données :

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                      /translation="MTQLQISLLLTATISLLHLVVATPYE
```