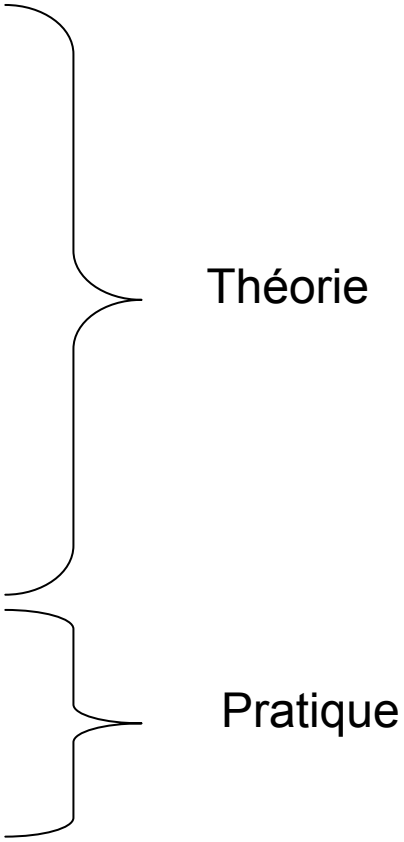
Sébastien Clément
Avril 2011

Plan du cours

1. Introduction
   - SGBD
   - SQL
2. Pourquoi utiliser un SGBD ?
3. Comment créer une base de données
   - besoins
   - modélisation
   - SQL
4. Interroger une base de données



SGBD

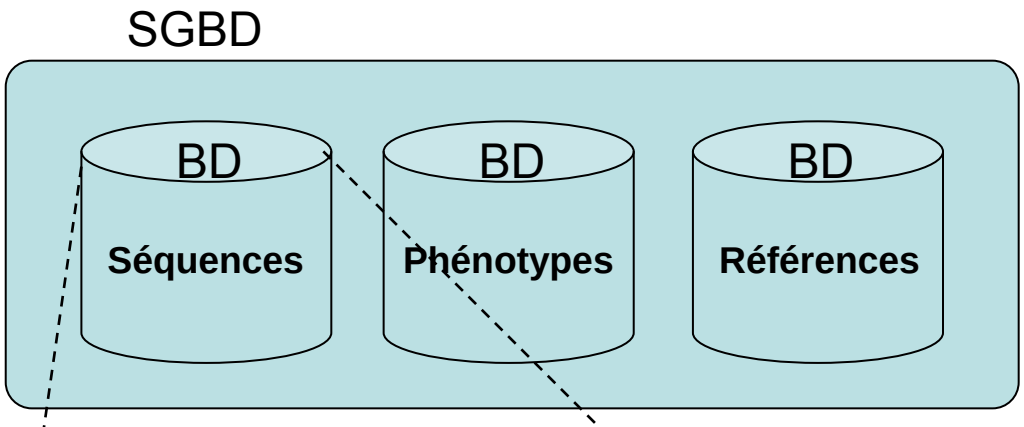
- Système de gestion de bases de données (anglais: *DBMS ou RDBMS*)

- Exemples:
  - Access
  - MS SQL Server
  - Oracle
  - MySQL
  - PostgreSQL

- Caractéristiques:
  - Langage commun: SQL
  - Système sécurisé
  - Opérations concurrentes
  - Intégrité
  - Catalogue
  - Sauvegarde et récupération
  - Vues

Création de bases de données SQL

Sébastien Clément
Avril 2011



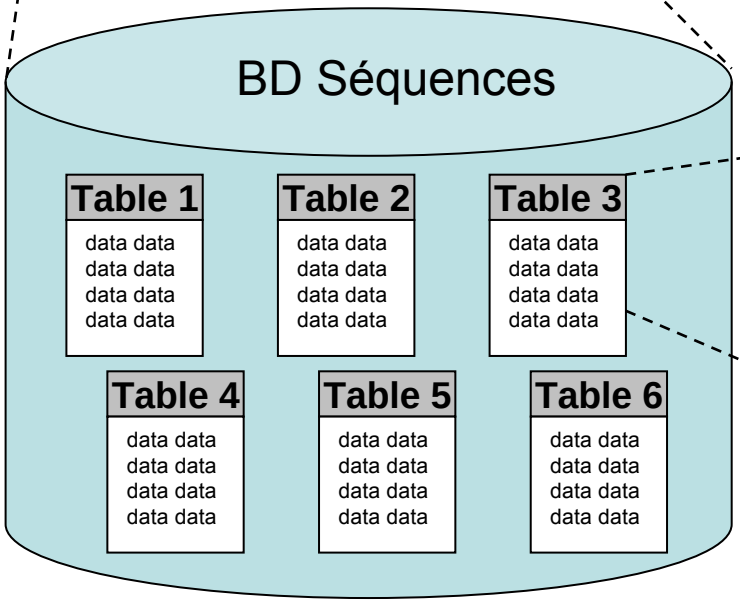


Table 3

id	name	sequence	sequence_position
1238	14528R2	CGAGCACTGAATAGCCTGGA	613
1353	4489F	TGAAGAGGCACAGACATTGC	28
1552	5963R2	TTCTTGCAACCATGGGAAAC	500
1944	8083 R1	GATCTCCAGAGAGGTTGCCC	439
12	5855-R1	CACTTAGGCAACATCCACTTACC	15
13	5855-F1	GCCTGCAAGAAGACCCTAGTT	586
14	5855-F2	CACAACCAGGACGTCAATTCT	756
15	5855-R2	CAGCTCCAATCTCACGAACTC	217
16	5879-R1	ACATCCTTTTGCACGCTCTT	54
17	5879-F1	TCATACCAACTGGATGGGGTA	517
18	5931-R1	TCAAATTGAGATGGCTTACAGG	51

Structure

Données

BD - exemples

- Banques de données biologiques publiques
  - Genbank
  - TAIR
  - Flybase
  - etc.

- De façon générale - tout ce qui gère des données:
  - Banques – ex: transactions aux guichets
  - Sites d'achat en ligne (Ebay, Amazon, Expedia, Hotwire, etc.)
  - GIS
  - etc.

SQL

- *Structured Query Language*
- Standard ISO
- Commun aux SGBD, mais certains ont des spécificités

- Plus que de la recherche de données:
  - DDL – Data Definition Language
  - DML – Data Manipulation Language

- Avantages:
  - Simple
  - Rapide
  - Non-procédural

- Exemple:

```
SELECT champ1,champ5 FROM ma_table;
```

Pourquoi utiliser un SGBD ?

	SGBD	Fichers à plat, Excel
+	•Données regroupées •Accès concurrent •Accès sécurisé •Sauvegarde, restauration •Contrôle d'intégrité (validation) •Mise-à-jour des données •Recherche facile (SQL) et rapide (index) •Vues	•Rapide à mettre sur pied •Facile d'utilisation (Excel)
-	•Plus long à monter •Doit être bien planifié •Performance de calcul •Taille	•Capacité limitée (Excel) •Aucun accès concurrent •Aucun contrôle d'intégrité •Accès non-sécurisé •Redondance et multiples-versions des données •Jointures ???

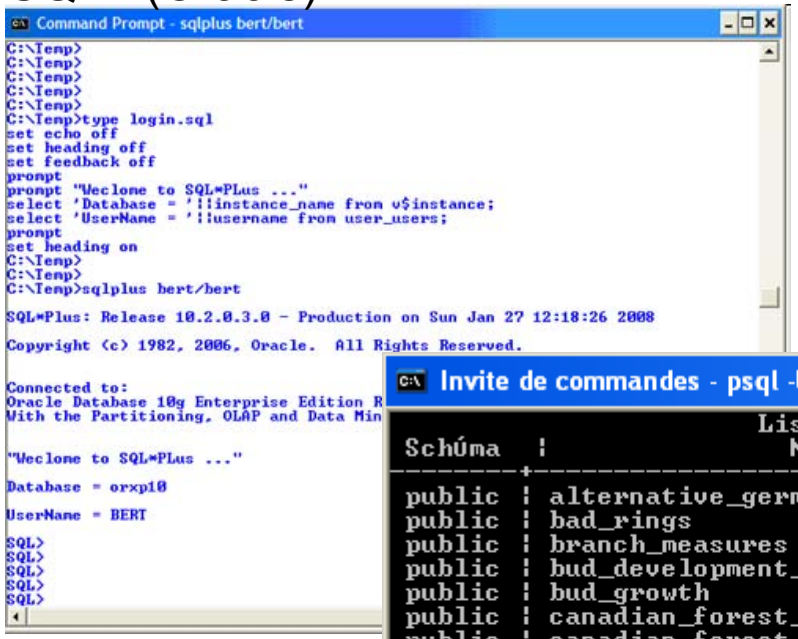
Comment choisir un SGBD ?

	Propriétaires	**Libres**
+	•Meilleur support technique* •Documentation	•Gratuit •Souvent aussi performant
-	•Prix** •Lourdeur	•Support technique souvent déficient (forums…)
	• MS Access • MS SQL Server • Oracle • Sybase	•MySQL •PostgreSQL •SQLite

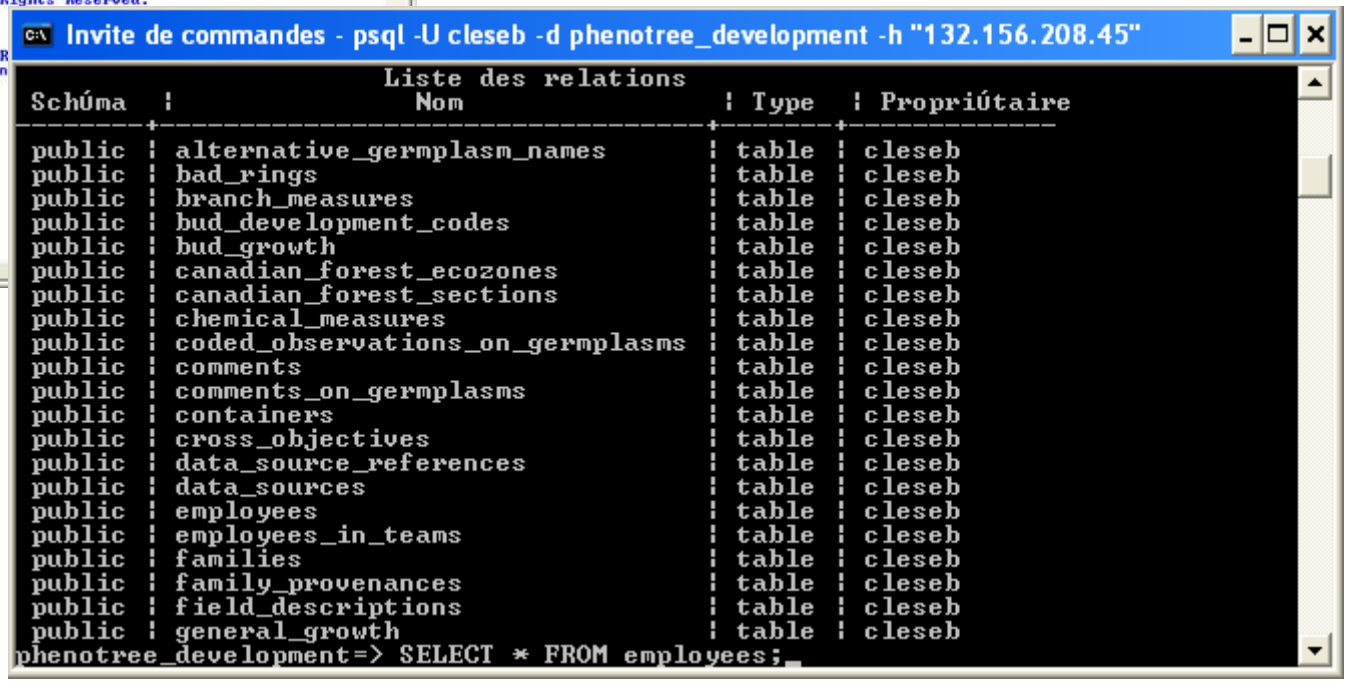
*Souvent coûts en sus du coût d'achat
**SQL Server: 9 500 à 36 700 $ CAN (selon version), license de 2 ans

Interfaces-utilisateur: ligne de commande (SQL) ou interface graphique ?

SQL+ (Oracle)



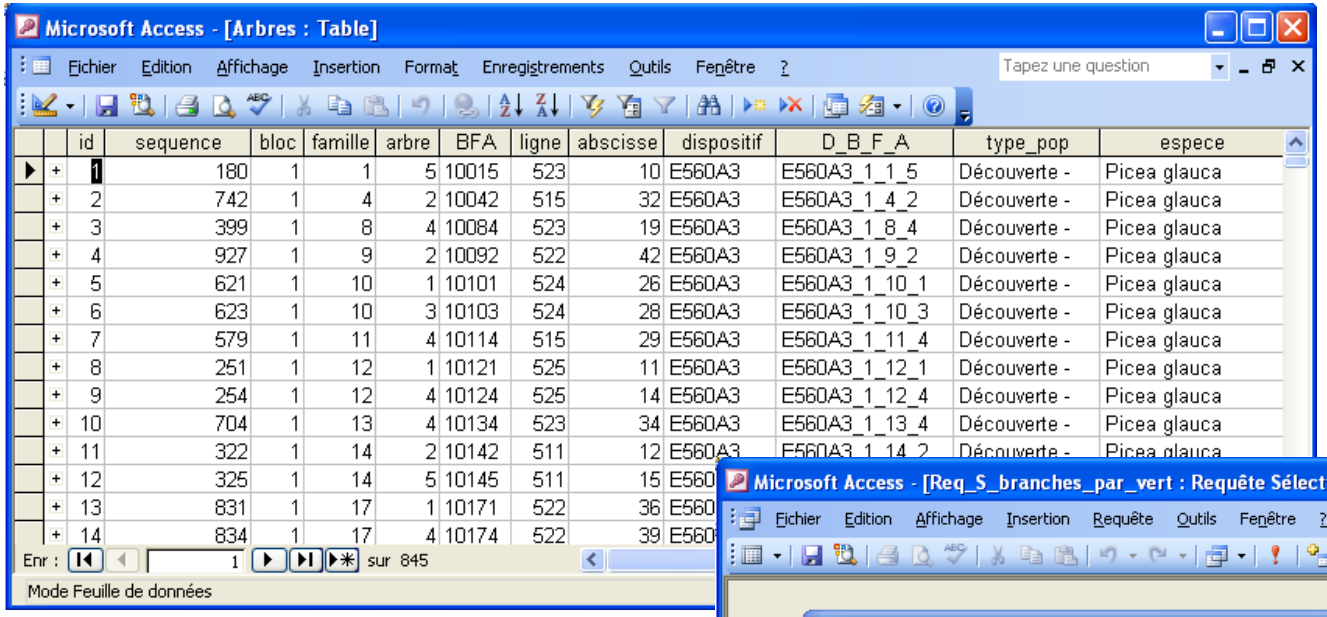
psql (PostgreSQL)

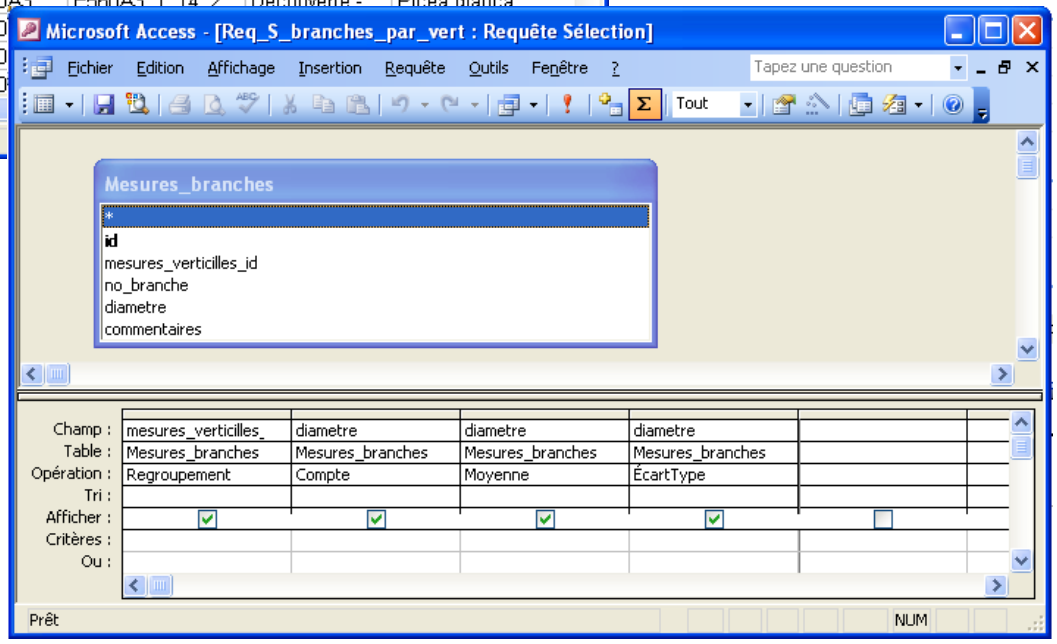


Sébastien Clément
Avril 2011

Interfaces-utilisateur: ligne de commande (SQL) ou interface graphique ?

Access – formulaire d'édition des enregistrements





Access – constructeur de requêtes

Création de bases de données SQL

Sébastien Clément
Avril 2011

Créer une base de données: besoins, modélisation et création SQL

Besoins de l'équipe

- Quel est le but ?
- Quelles données ?
- Utilisation concurrente ?
- Court-terme vs long-terme ?
- Rôles des utilisateurs ?
  - administrateur de la bd (DBA)
  - administrateur des données (DA)
  - utilisateur en lecture seule

Créer une base de données: besoins, modélisation et création SQL

La modélisation des données

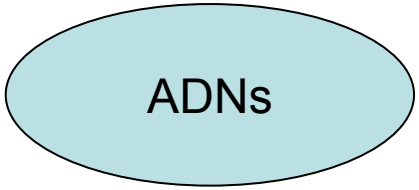
Modèle de données:

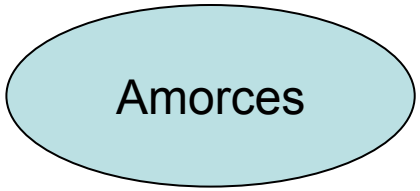
« ***Description des données (1), des liens entre elles (2) et des contraintes sur les données (3), dans une organisation.*** »

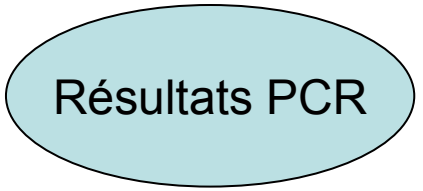
1. Description des données

Mon équipe de recherche a:

- ***ADN (cDNA, gDNA)***
- ***amorces***
- ***résultats PCR***

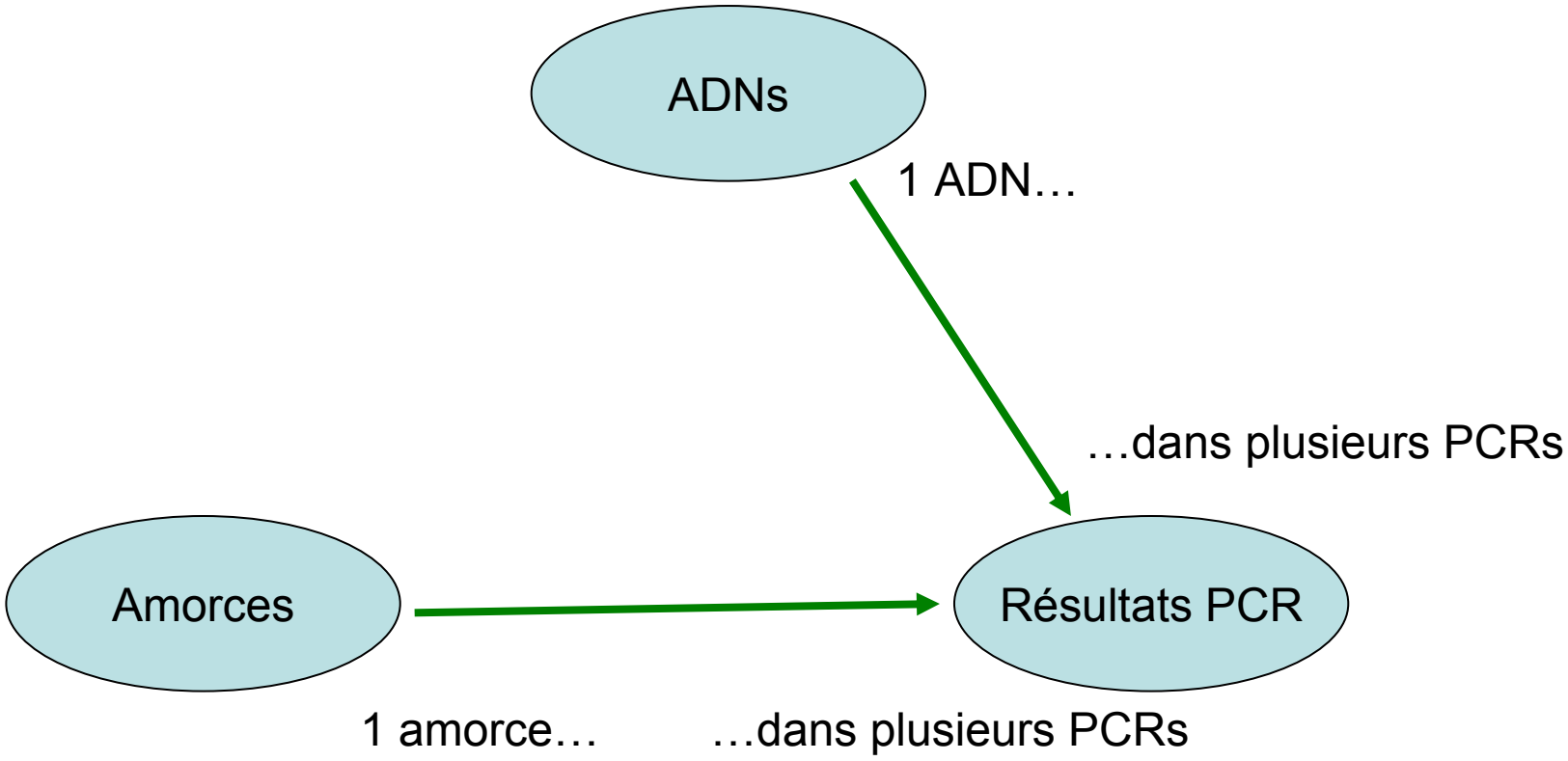




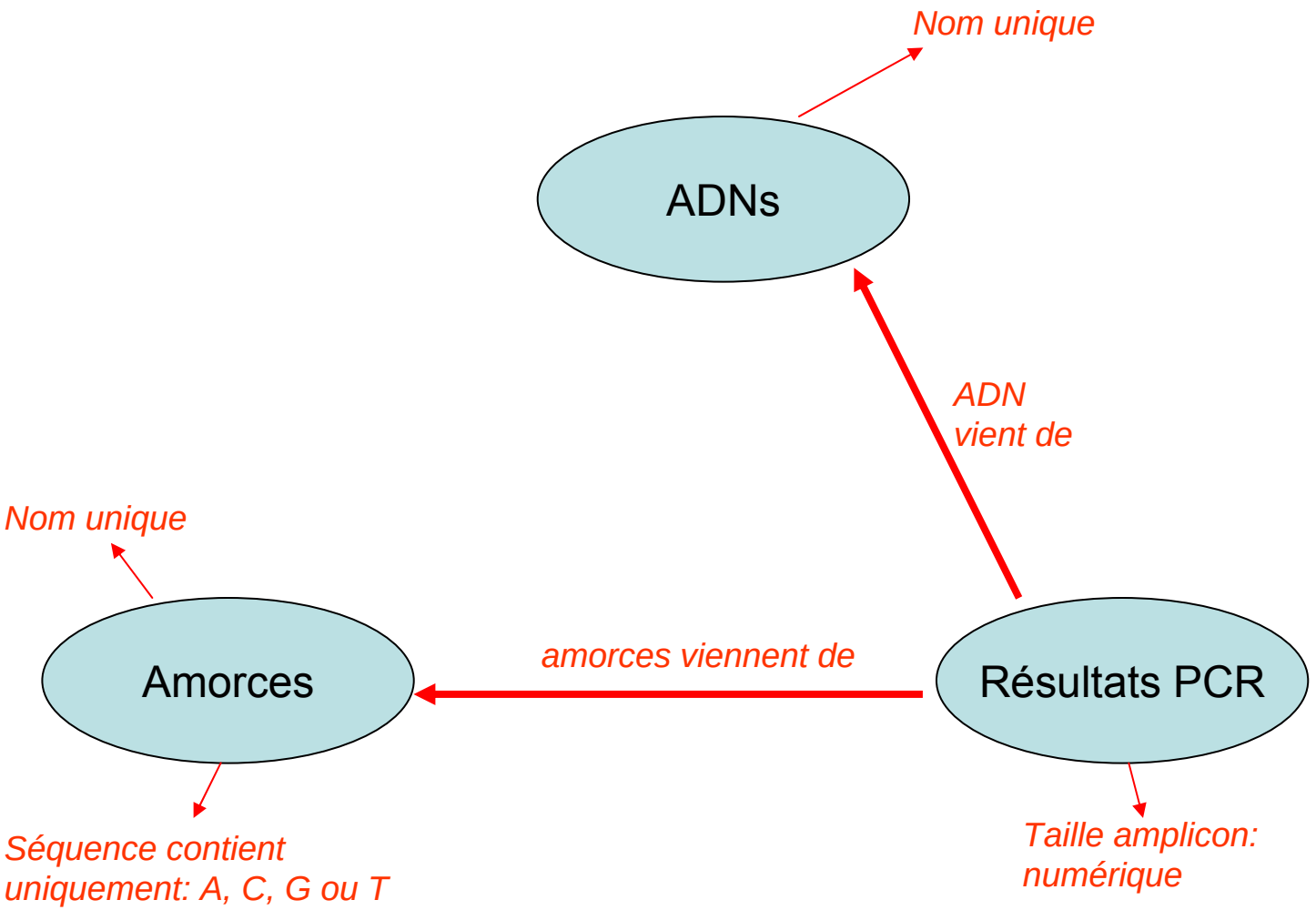


2. Liens entre les données

- ***Les amorces et les ADN sont mis en commun pour faire un PCR***



3. Contraintes sur les données



Sébastien Clément
Avril 2011

1. Description des données

Mon équipe de recherche a:

- ***ADN***
- ***amorces***
- ***résultats PCR***

ADNs

nom	espèce	type
at_312	A. thaliana	génomique
pt_221s	P. taeda	cDNA
pg_S352	P. glauca	cDNA
ADN_K33	P. aeruginosa	plasmidique
ce_ttY14	C. elegans	cDNA

Amorces

nom	séquence	sens
at39234F	ACACACAACAACACAATTCG	F
at39234R	CATACATATCATACTATCAACTA	R
pt00324F	GCCTGCAAGAAGACCCTAGTT	F
pt00324R	CACTTAGGCAACATCCACTTACC	R

PCRs

adn_nom	amorce_f	amorce_r	programme	melange	taille_amplicon
at_312	at39234F	at39234R	Prog0077	Mix321	
at_312	at39234F	at39234R	Prog0002	Mix321	1400
ADN_K33	K33_001F	K33_001R	Prog0012	Mix217	
ADN_K33	K33_001F	K33_001R	Prog0012	Mix218	450

2. Liens entre les données

Les amorces et les ADN sont mis en commun pour faire un PCR

ADNs

nom	espèce	type
at_312	A. thaliana	génomique
pt_221s	P. taeda	cDNA
pg_S352	P. glauca	cDNA
ADN_K33	P. aeruginosa	plasmidique
ce_ttY14	C. elegans	cDNA

1

∞

Amorces

nom	séquence	sens
at39234F	ACACACAACAACACAATTCG	F
at39234R	CATACATATCATACTATCAACTA	R
pt00324F	GCCTGCAAGAAGACCCTAGTT	F
pt00324R	CACTTAGGCAACATCCACTTACC	R

1

∞ ∞

PCRs

adn_nom	amorce_f	amorce_r	programme	melange	taille_amplicon
at_312	at39234F	at39234R	Prog0077	Mix321	
at_312	at39234F	at39234R	Prog0002	Mix321	1400
ADN_K33	K33_001F	K33_001R	Prog0012	Mix217	
ADN_K33	K33_001F	K33_001R	Prog0012	Mix218	450

3. Contraintes sur les données

ADNs

nom	espèce	type
at_312	A. thaliana	génomique
pt_221s	P. taeda	cDNA
pg_S352	P. glauca	cDNA
ADN_K33	P. aeruginosa	plasmidique
ce_ttY14	C. elegans	cDNA

Unique

Amorces

nom	séquence	sens
at39234F	ACACACAACAACACAATTCG	F
at39234R	CATACATATCATACTATCAACTA	R
pt00324F	GCCTGCAAGAAGACCCTAGTT	F
pt00324R	CACTTAGGCAACATCCACTTACC	R

A, C, G ou T

Doit provenir de

PCRs

adn_nom	amorce_f	amorce_r	programme	melange	taille_amplicon
at_312	at39234F	at39234R	Prog0077	Mix321	
at_312	at39234F	at39234R	Prog0002	Mix321	1400
ADN_K33	K33_001F	K33_001R	Prog0012	Mix217	
ADN_K33	K33_001F	K33_001R	Prog0012	Mix218	450

Numérique

Sébastien Clément
Avril 2011

3. Contraintes sur les données

Primaire (PK):
identifie une ligne de façon unique

Étrangère (FK):
réfère à la clé primaire d'une autre table

ADNs

nom	espèce	type
at_312	A. thaliana	génomique
pt_221s	P. taeda	cDNA
pg_S352	P. glauca	cDNA
ADN_K33	P. aeruginosa	plasmidique
ce_ttY14	C. elegans	cDNA

Amorces

nom	séquence	sens
at39234F	ACACACAACAACACAATTCG	F
at39234R	CATACATATCATACTATCAACTA	R
pt00324F	GCCTGCAAGAAGACCCTAGTT	F
pt00324R	CACTTAGGCAACATCCACTTACC	R

?

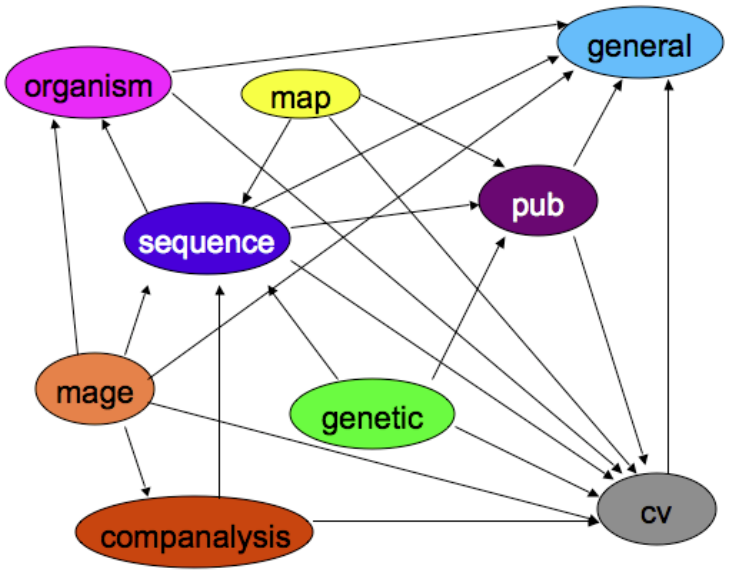
PCRs

id	adn_nom	amorce_f	amorce_r	programme	melange	taille_amplicon
1	at_312	at39234F	at39234R	Prog0077	Mix321	
2	at_312	at39234F	at39234R	Prog0002	Mix321	1400
3	ADN_K33	K33_001F	K33_001R	Prog0012	Mix217	
4	ADN_K33	K33_001F	K33_001R	Prog0012	Mix218	450

Schémas déjà existants sur le Web?

Génomique

Chado - http://gmod.org/wiki/Chado



BioSQL - http://www.biosql.org/wiki/Main_Page

+	•Générique •Plusieurs sont gratuits •Déjà utilisé et amélioré par d'autres •Interfaces graphiques
-	•Générique ! •Maîtrise du schéma: long •« Overkill » ?

Création de bases de données SQL

Sébastien Clément
Avril 2011

La création de la base de données en SQL

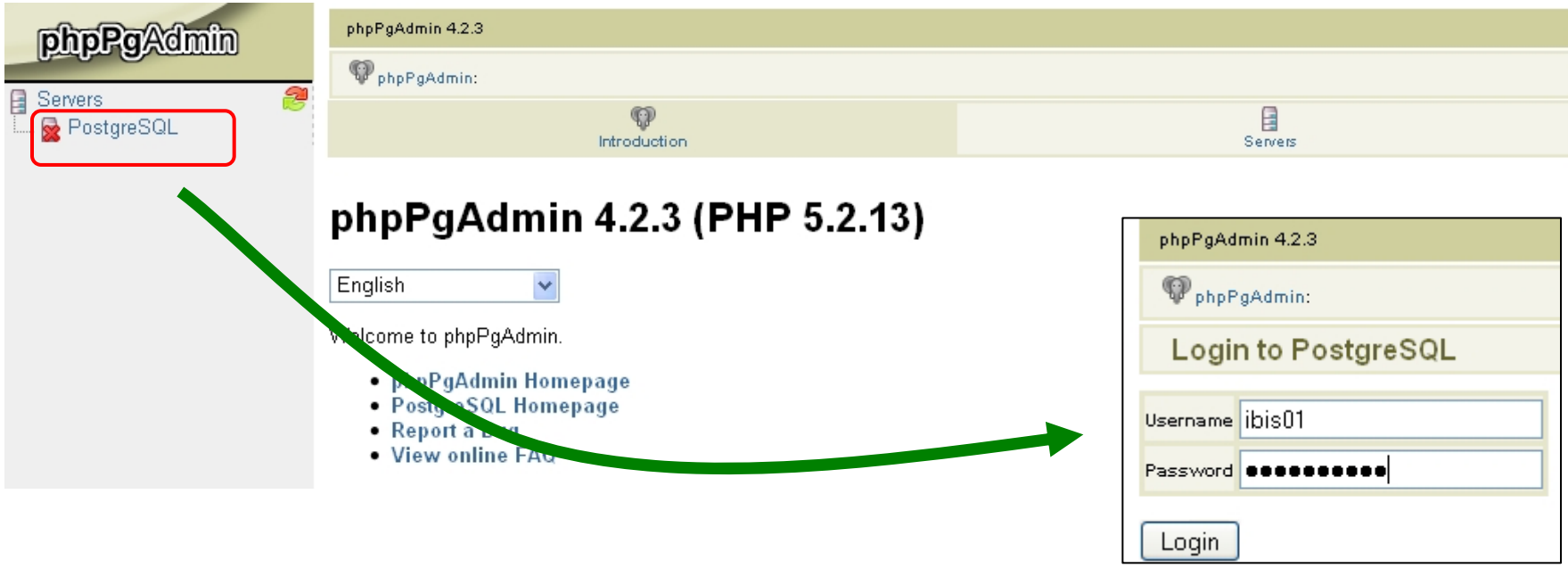
Serveur de bases de données

SGBD: **PostgreSQL**

Site: http://dbs.arborea.ulaval.ca/phppgadmin/

Interface Web: PhpPgAdmin

Base de données: **ibis_db**



Partie pratique 1

Créer une base de données

Sébastien Clément
Avril 2011

Interroger une base de données – requête SQL (DML)

À quoi servent les requêtes SQL ?

- A) Aller chercher de l'info de façon très précise (critères)
- B) Joindre de l'info répartie dans plusieurs tables
- C) Effectuer des calculs par ligne, ou sur plusieurs lignes

Avantages:

- Langage simple
- Dynamique (mise-à-jour)
- Soumis aux privilèges d'accès utilisateurs

Sébastien Clément
Avril 2011

Interroger une base de données – requête SQL (DML)

Exemple A: Aller chercher de l'info de façon très précise (critères)

Table amorces

nom	sequence	sens	adn_gabarit	pos_vs_gabarit	date_creation
at39234F	ACACACAACAACACAATTCG	F	at_312	43	2011-02-12
at39234R	CATACATATCATACTATCAACTA	R	at_312	512	2011-02-12
pt00324F	GCCTGCAAGAAGACCCTAGTT	F	pt_221s	52	2011-03-05
pt00324R	CACTTAGGCAACATCCACTTACC	R	pt_221s	744	2011-03-05
K33_001F	CCAGATCTACTACATACTTACTAT	F	ADN_K33	23	2010-09-07
K33_001R	CAACATCTATCTACTATCTAT	R	ADN_K33	455	2010-09-07
K33_002F	CACACACACAATACGACACAT	F	ADN_K33	433	2010-09-07
K33_002R	CGGCGATTATTCGATTACGATCT	R	ADN_K33	971	2010-09-07

Sélection des colonnes

Sélection des lignes

nom	date_creation
at39234F	2011-02-12
at39234R	2011-02-12
pt00324F	2011-03-05
pt00324R	2011-03-05
K33_001F	2010-09-07
K33_001R	2010-09-07
K33_002F	2010-09-07
K33_002R	2010-09-07

nom	sequence	sens	adn_gabarit	pos_vs_gabarit	date_creation
at39234F	ACACACAACAACACAATTCG	F	at_312	43	2011-02-12
pt00324F	GCCTGCAAGAAGACCCTAGTT	F	pt_221s	52	2011-03-05
K33_001F	CCAGATCTACTACATACTTACTAT	F	ADN_K33	23	2010-09-07
K33_002F	CACACACACAATACGACACAT	F	ADN_K33	433	2010-09-07

Sébastien Clément
Avril 2011

Interroger une base de données – requête SQL (DML)

Exemple B: Joindre de l'info répartie dans plusieurs tables

Table ADNs

nom	espece	type
at_312	A. thaliana	Génomique
pt_221s	P. taeda	cDNA
pg_S352	P. glauca	cDNA
ADN_K33	P. aeruginosa	Plasmidique
ce_ttY14	C. elegans	cDNA

Table amorces

nom	sequence	sens	adn_gabarit	pos_vs_gabarit	date_creation
at39234F	ACACACAACAACACAATTCG	F	at_312	43	2011-02-12
at39234R	CATACATATCATACTATCAACTA	R	at_312	512	2011-02-12
pt00324F	GCCTGCAAGAAGACCCTAGTT	F	pt_221s	52	2011-03-05
pt00324R	CACTTAGGCAACATCCACTTACC	R	pt_221s	744	2011-03-05
K33_001F	CCAGATCTACTACATACTTACTAT	F	ADN_K33	23	2010-09-07
K33_001R	CAACATCTATCTACTATCTAT	R	ADN_K33	455	2010-09-07
K33_002F	CACACACACAATACGACACAT	F	ADN_K33	433	2010-09-07
K33_002R	CGGCGATTATTCGATTACGATCT	R	ADN_K33	971	2010-09-07

ADNs et amorces correspondantes

nom	espece	type	nom	sequence
at_312	A. thaliana	Génomique	at39234R	CATACATATCATACTATCAACTA
at_312	A. thaliana	Génomique	at39234F	ACACACAACAACACAATTCG
pt_221s	P. taeda	cDNA	pt00324R	CACTTAGGCAACATCCACTTACC
pt_221s	P. taeda	cDNA	pt00324F	GCCTGCAAGAAGACCCTAGTT
pg_S352	P. glauca	cDNA	*NULL*	*NULL*
ADN_K33	P. aeruginosa	Plasmidique	K33_002R	CGGCGATTATTCGATTACGATCT
ADN_K33	P. aeruginosa	Plasmidique	K33_002F	CACACACACAATACGACACAT
ADN_K33	P. aeruginosa	Plasmidique	K33_001R	CAACATCTATCTACTATCTAT
ADN_K33	P. aeruginosa	Plasmidique	K33_001F	CCAGATCTACTACATACTTACTAT
ce_ttY14	C. elegans	cDNA	*NULL*	*NULL*

Interroger une base de données – requête SQL (DML)

Exemple C: Effectuer des calculs par ligne, ou sur plusieurs lignes

Table amorces

nom	sequence	sens	adn_gabarit	pos_vs_gabarit	date_creation
at39234F	ACACACAACAACACAATTCG	F	at_312	43	2011-02-12
at39234R	CATACATATCATACTATCAACTA	R	at_312	512	2011-02-12
pt00324F	GCCTGCAAGAAGACCCTAGTT	F	pt_221s	52	2011-03-05
pt00324R	CACTTAGGCAACATCCACTTACC	R	pt_221s	744	2011-03-05
K33_001F	CCAGATCTACTACATACTTACTAT	F	ADN_K33	23	2010-09-07
K33_001R	CAACATCTATCTACTATCTAT	R	ADN_K33	455	2010-09-07
K33_002F	CACACACACAATACGACACAT	F	ADN_K33	433	2010-09-07
K33_002R	CGGCGATTATTCGATTACGATCT	R	ADN_K33	971	2010-09-07

Calcul par ligne

Calcul sur plusieurs lignes

nom	sequence	pct_gc	length
at39234F	ACACACAACAACACAATTCG	40.0	20
at39234R	CATACATATCATACTATCAACTA	26.1	23
pt00324F	GCCTGCAAGAAGACCCTAGTT	52.4	21
pt00324R	CACTTAGGCAACATCCACTTACC	47.8	23
K33_001F	CCAGATCTACTACATACTTACTAT	33.3	24
K33_001R	CAACATCTATCTACTATCTAT	28.6	21
K33_002F	CACACACACAATACGACACAT	42.9	21
K33_002R	CGGCGATTATTCGATTACGATCT	43.5	23

Champs calculés

sens	nb_amorces	avg_pos	min_pos	max_pos
F	4	138	23	433
R	4	671	455	971

Partie pratique 2

Interroger une base de données