GBds
Genomic Breeding decision support

Carole.Bernon@irit.fr

Journées Big Data – UT3 – 17/11/15

Genomic Breeding decision support

- **Partenaires**
  - RAGT 2n, coordinateur (semencier)
  - Euralis (semencier)
  - Météo-France
  - Upetec (JEI, à l'origine start-up de SMAC)
  - IRIT : équipes APO & SMAC
- **Financeurs**
  - Projet FUI
    - BPI France (Oseo)
    - Régions Midi-Pyrénées & Aquitaine
    - Conseil général de l'Aveyron, Grand Rodez
    - Union Européenne via fonds FEDER
- **Label**
  - AgriSudOuest Innovation
- **Durée**
  - 3 ans : mai 2012-juin 2015

Objectif du projet

- **Cadre**
  - Développer des outils de sélection génomique novateurs pour créer des variétés de maïs performantes et respectueuses de l'environnement
  - Modèles mathématiques encore insuffisants
- **Objectifs**
  - Outil permettant de prédire la valeur phénotypique d'un hybride à partir des caractéristiques génomiques de leurs lignées (parents)
  - Identifier les hybrides les mieux adaptés à des conditions environnementales données
- **Cible**
  - Sélectionneurs de variétés
    - Aider à faire le bon choix en amont d'essais au champ coûteux

Verrou scientifique

- **Les caractéristiques d'une plante s'expliquent**
  - à 30% par la génétique (effet G)
  - à 30% par l'environnement (effet E)
  - à 40% par l'interaction génétique/environnement (effet GxE)
- **Mais l'effet GxE n'est actuellement que peu étudié**
  - Explosion des données dès que l'on considère l'environnement
  - Mauvaise qualité des données relevées (erreurs, oublis…)
  - Méthodes de prédiction classiques (statistiques) peu génériques
    - Issues du monde animal et négligeant l'effet E (animaux moins sensibles à leur environnement)
    - Traitent les données *a priori* (comblent les manques, éliminent les données jugées non cohérentes…)
    - Solutions *ad hoc*

Des données…

- **Hétérogènes : discrètes et/ou continues**
  - Génomiques (fournies par les semenciers)
    - 60K marqueurs (SNP) : des individus testés en champ ou de leurs parents
  - Environnementales (fournies par Météo-France)
    - Météorologiques : 5 variables journalières (température, pluviométrie…)
    - Pédologiques : 2 niveaux de réserves utiles en eau du sol
  - Agronomiques (fournies par les semenciers)
    - 20 variables relevées au champ : rendement, verse, humidité du grain, date de floraison, parasitage…
  - Variables composites délimitant les différentes phases du maïs
    - Env. 100, (pré)calculées, pour chaque individu
- **Massives**
  - Concernant les années 1997 à 2014
  - BD de quelques dizaines Gigas pour l'instant…
- **Bruitées**
- **Manquantes**

Deux types d'algorithmes de prédiction

- **Approche « classique » (équipe APO) = référence**
  - Modèle « *Completely Multiplicative Model* » (COMM)
    - Sépare G et E
  - Modèle « *Ridge Regression* »
    - Calcule la valeur moyenne génétique pour calibrer le MLM
  - Modèle linéaire mixte (MLM) de type BLUP
    - Corrélation génétique-phénotype
  - Prétraitement des données
  - Ne peut prédire pour un environnement « non rencontré »
- **Approche « nouvelle » (UPETEC & équipe SMAC)**
  - Reposant sur les AMAS (*Adaptive Multi-Agent Systems*)
  - Prétraitement des données inutile
  - Traiter toutes les données de la même manière
  - Prédire pour un environnement « non rencontré »
- **Objectif : comparer les 2 approches**

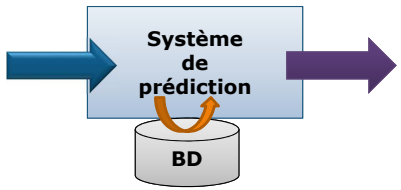
Approche par les AMAS

- **Vu comme un problème d'optimisation combinatoire**
  - AMAS = heuristique de recherche
- **Principe**

Donnée	D120	D230	D450	D3	. . . .	D2010
Rendement	120	152	154	145	. . . .	165

Valeur phénotypique + variables (G, E) avec des contraintes (ou non) (Requête sélectionneur)

Valeur recherchée	Rendement
Marqueur 1	AA
Marqueur 3	GT
Pluie	20



Données
(1 donnée = valeur phénotypique + vecteur des valeurs des variables génomiques & environnementales associées)

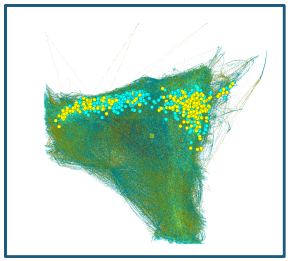
Ensemble de p données constituant les meilleurs « prédicteurs »

Valeur phénotypique prédite déduite

156

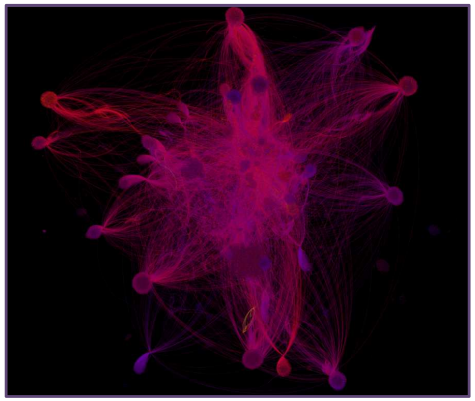
Donnée	Rendement	Marqueur 1	Marqueur 2	Marqueur 3	Température	Pluie	Etc.
D10	164	AA	AT	GA	22	10	. . . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . . .
D347	155	AA	TA	GC	21	15	. . . . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . . .

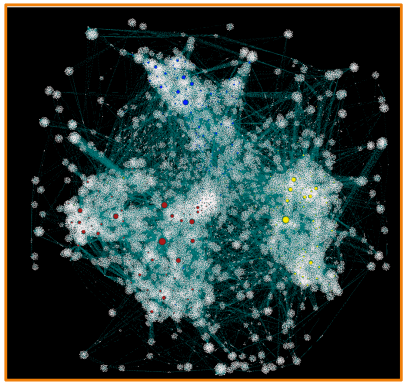
Outils développés (1a/2)



- **Exploration de la BD**

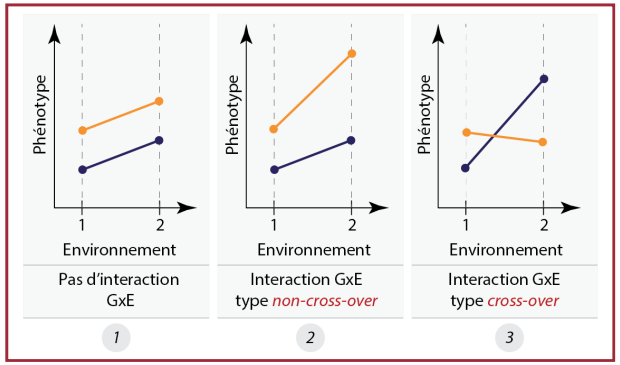
Outils développés (1b/2)





- Graphe de voisinage génétique

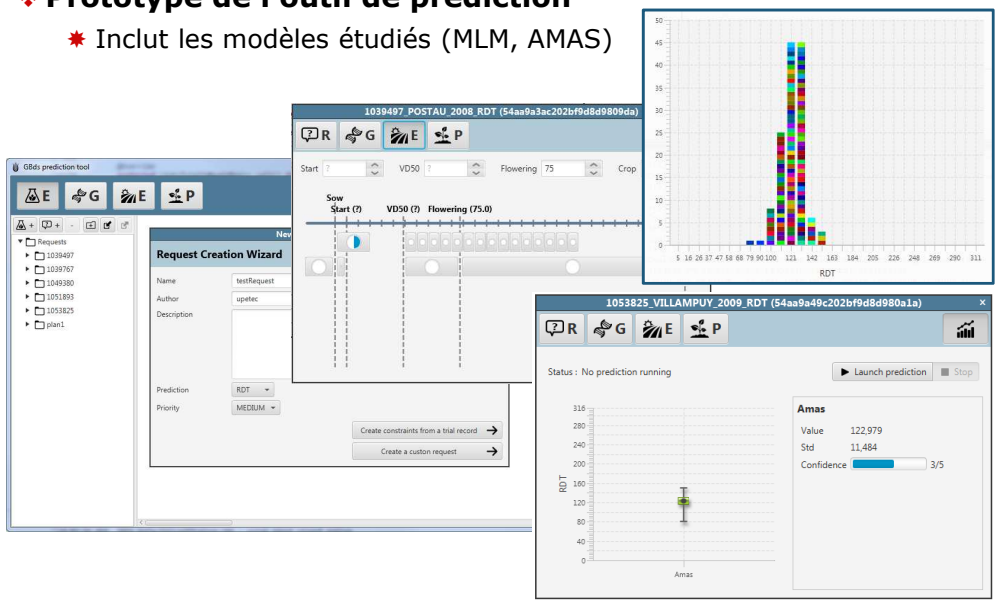
Outils développés (1c/2)



- Générateur synthétique de données

Outils développés (2/2)

- Prototype de l'outil de prédiction
  - Inclut les modèles étudiés (MLM, AMAS)



Protocole de validation

- 4 types de tests
  - Menés en « Leave one out »
  - Sans accès à « Individu + Environnement (lieu + date de semis) » de la requête
- À partir de données
  - Expérimentales
  - Simulées
- Avec / Sans contraintes environnementales
- Quelques milliers de tests
- Comparaison sur 285 tests (IIEC)

	Individus connus	Individus inconnus
Environnements connus	**Tests ICEC** posent un problème trivial consistant à récupérer l'information dans la BD. *Le système réalise une prédiction très proche de la réalité en cherchant les éléments proches de la cible.*	**Tests IIEC** menés en Leave-One-Out. *le système est autorisé à utiliser des informations en provenance d'autres individus du même essai*
Environnements inconnus	**Tests ICEI** à explorer ultérieurement	**Tests IIEI**

Problèmes rencontrés

- **Principalement liés à la qualité des données fournies**
  - Spécificités propres à chaque semencier
  - Variétés et précocités des maïs parfois mélangées
  - Erreurs de relevés sur certains phénotypes
  - …
- **Biais induit par la pratique des sélectionneurs**
  - Corrélation entre localisation des tests et génétique des individus testés
- **Remise en cause d'une partie des tests**
- **Difficulté de (auto)calibrer l'apprentissage**

Bilan

- **Approches complémentaires**
  - MLM : Données génotypées ET Environnements connus
  - AMAS : Données non génotypées OU Environnement « inconnu »
- **AMAS plus intéressant**
  - Pour l'effet G ou E seuls, parfois
  - Pour supporter mieux les données manquantes et le bruit
- **Méthode de prédiction générique**
  - Appliquée à d'autres types de données (réseaux sociaux)
- **Beaucoup de potentialités pour l'outil conçu**
  - Extension à d'autres variétés (tournesol…)
  - Améliorer la qualité des données en identifiant celles aberrantes en BD
  - Extension à d'autres domaines où prédire un caractère ou un comportement est utile (pharmacologie, énergie, systèmes ambiants, suivi de patients au domicile…)

Merci pour votre attention

GBds