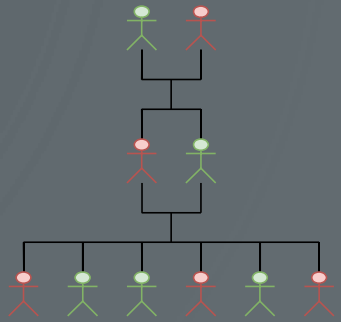
MIA TOULOUSE

Arbres de Jonction Hiérarchiques

Pour l'inférence de génotypes dans les pedigrees complexes

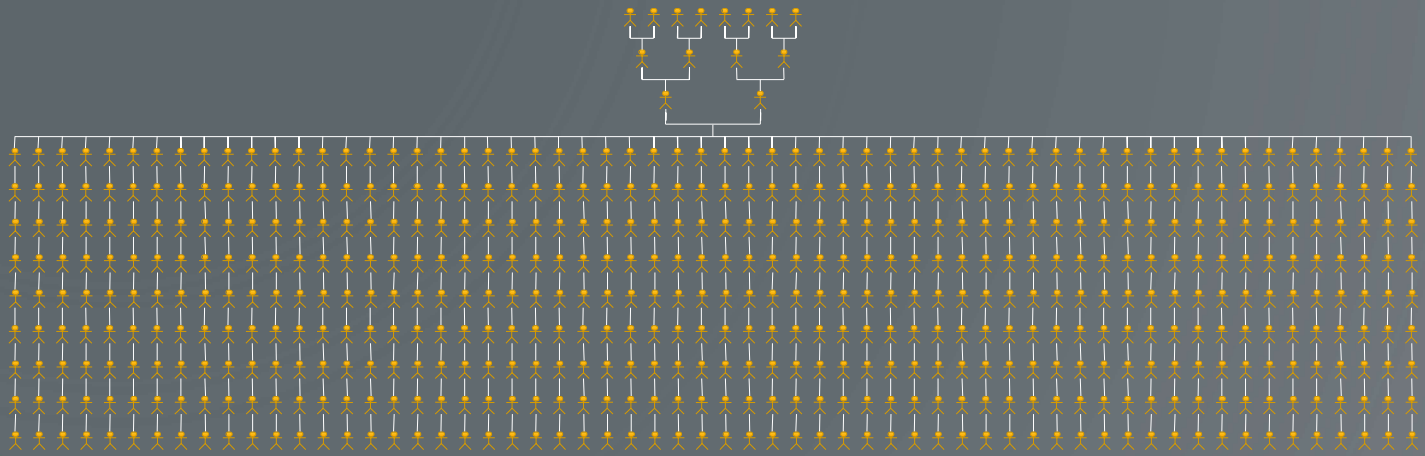
Génétique quantitative et pedigrees

pedigrees



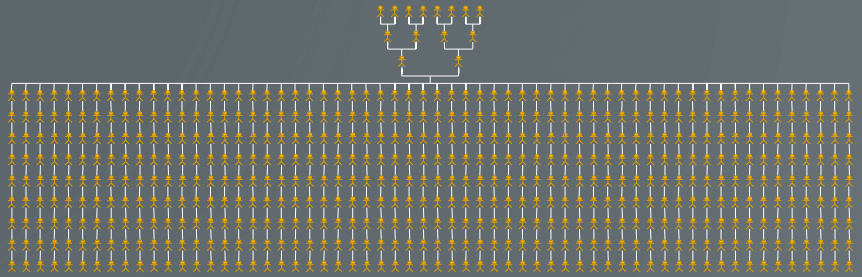
Génétique quantitative et pedigrees

pedigrees



Génétique quantitative et pedigrees

pedigrees

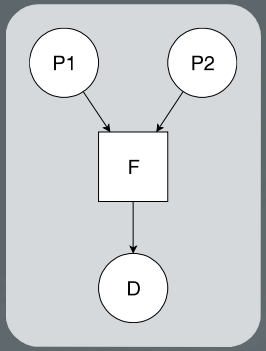


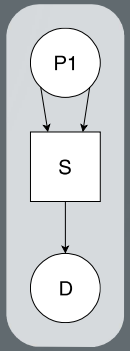
- Succession de croisements entre individus ou d'auto-fécondations (plantes)
- Observations possibles sur certains individus
- Questions posées sur certains individus
  - Inférence de génotypes

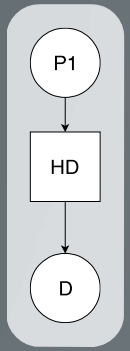
  ⇒ Quelles sont les probabilités que tel individu hérite de tel ancêtre ?
- Applications en cartographie génétique et analyse QTL

Génétique quantitative et pedigrees

Structure particulière d'un pedigree

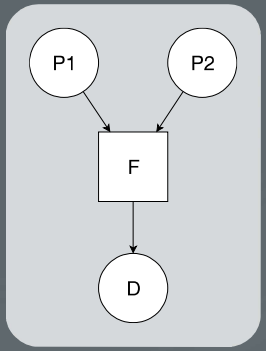




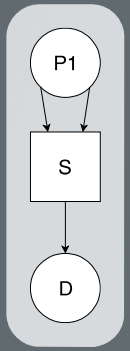


Génétique quantitative et pedigrees

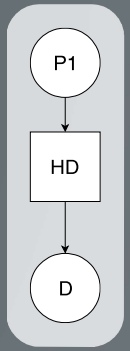
Structure particulière d'un pedigree



⇒Hétérogamie
(**2** gamètes)



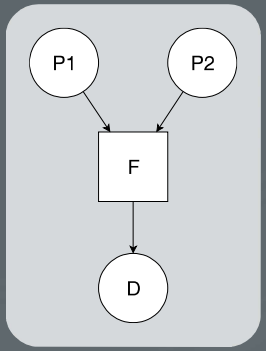
⇒Autofécondation
(**2** gamètes)



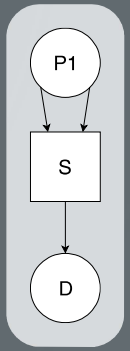
⇒Haploïde doublé
(**1** gamète)

Génétique quantitative et pedigrees

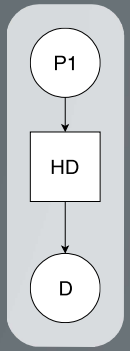
Structure particulière d'un pedigree



⇒Hétérogamie
(**2** gamètes)



⇒Autofécondation
(**2** gamètes)



⇒Haploïde doublé
(**1** gamète)

- Degré entrant limité (0, 1, 2)
- Degré sortant non borné

Génétique quantitative et pedigrees

Domaine des variables

- Les variables sont les individus du pedigree. On s'intéresse à l'origine génétique des individus.
- Comme il y a un brin chromosomique hérité de chaque parent, le domaine est au moins

$$\text{ancêtres} \times \text{ancêtres}$$

- Comme les observations sont souvent des allèles (*e.g.* SNP), le domaine devient :

$$\text{ancêtres} \times \text{allèles} \times \text{ancêtres} \times \text{allèles}$$

Si l'on considère un pedigree MAGIC à 8 parents et des observations de SNPs bi-alléliques, le cardinal est

$$8^2 * 2^2 = 256$$

⇒ Le coût d'ajout d'une variable à une clique devient rapidement prohibitif !
Les tables sont néanmoins très creuses.

- … Des considérations sur la séparation en plusieurs variables ne font que déplacer le problème…

Représentation en graphe de facteurs

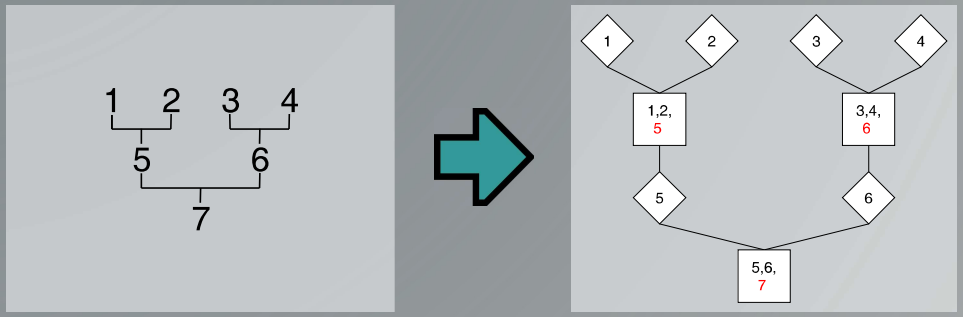
Un individu, un croisement, un facteur.

- Il s'agit de représenter le mélange des patrimoines génétiques.
- Les ancêtres dans le pedigree sont des tables mono-variables…
  Origine génétique fixée, allèles variables.
- Les croisements sont des tables à deux ou trois variables (très creuses).

Représentation en graphe de facteurs

Un individu, un croisement, un facteur.

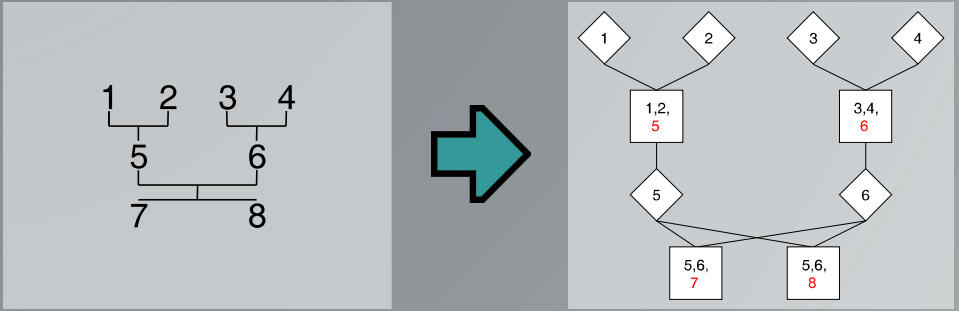
- Il s'agit de représenter le mélange des patrimoines génétiques.
- Les ancêtres dans le pedigree sont des tables mono-variables…
  Origine génétique fixée, allèles variables.
- Les croisements sont des tables à deux ou trois variables (très creuses).



Représentation en graphe de facteurs

Un individu, un croisement, un facteur.

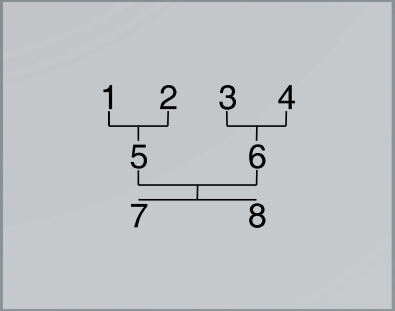
- Il s'agit de représenter le mélange des patrimoines génétiques.
- Les ancêtres dans le pedigree sont des tables mono-variables…
  Origine génétique fixée, allèles variables.
- Les croisements sont des tables à deux ou trois variables (très creuses).

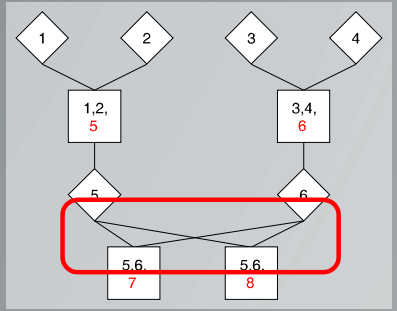


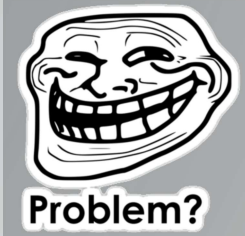
Représentation en graphe de facteurs

Un individu, un croisement, un facteur.

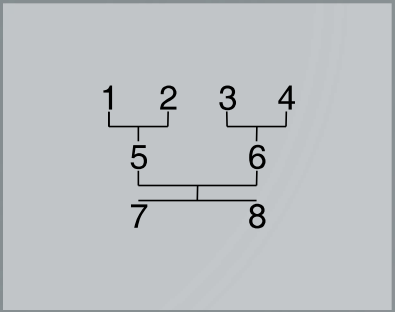
- Il s'agit de représenter le mélange des patrimoines génétiques.
- Les ancêtres dans le pedigree sont des tables mono-variables…
  Origine génétique fixée, allèles variables.
- Les croisements sont des tables à deux ou trois variables (très creuses).

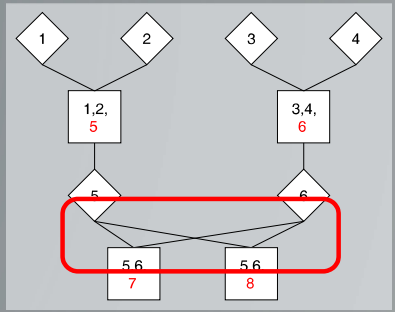


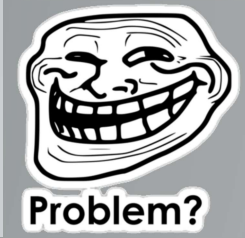




Représentation en graphe de facteurs







Cliques en folie

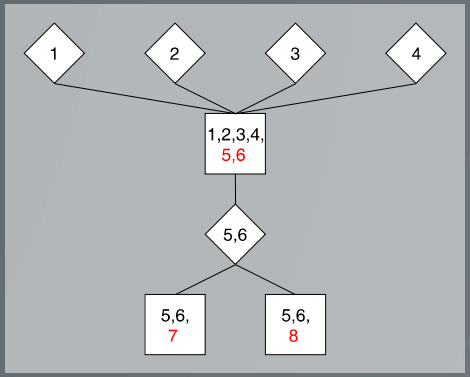
- Dès qu'il y a au moins deux descendants par paire de parents, des cycles apparaissent.
- Or, les pedigrees sont typiquement très larges.
- Une élimination de variables classique aboutit à mettre toute la famille dans la même clique.
  - ⇒ Beaucoup trop coûteux !
  - ⇒ On considère des probabilités jointes entre individus indépendants sachant leurs parents.

Représentation en graphe de facteurs

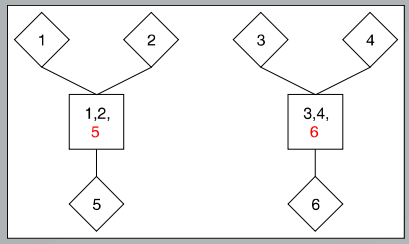
Vers une représentation plus saine

- Intuitivement, on voudrait regrouper comme ceci:
- Mais 5 et 6 sont indépendants !
- On voit un facteur comme une table de probabilité jointe, mais dans le cas présent

$$P(5,6 \mid 1,2,3,4) = P(5 \mid 1,2)\ P(6 \mid 3,4)$$



- Il n'y a pas besoin de le représenter sous forme de table. C'est en fait :



- Nous allons donc considérer un graphe de facteur où certains nœuds facteurs sont eux-mêmes des graphes.

Arbres de jonction hiérarchiques

Principe

- Un AJH est un graphe bipartite composé de facteurs et d'interfaces.
- Un facteur peut être une table de probabilités jointes ou un sous-AJH.
  - ⇒ Deux méthodes de calcul des beliefs
  - ⇒ Même comportement vis-à-vis de ses voisins
- Chaque niveau est un arbre ou une forêt.

Avantages

- Représentation compacte de la loi jointe du réseau.
- Pas de création artificielle de dépendances entre variables.
- [CONJECTURE] Cette méthode peut s'appliquer à n'importe quelle structure.

Caveat

Spécificités de notre contexte

- Chaque facteur représentant un croisement, il "produit" exactement une variable. Chaque variable est "produite" par exactement un facteur.
- On a une équivalence conceptuelle individu — variable — facteur.
- Le **rang** de chaque variable est bien défini et connu.

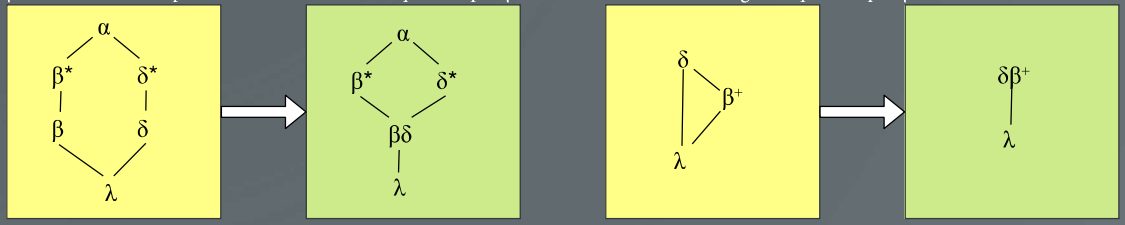
$$R(i) = 1 + \max(R(p^1{}_i), R(p^2{}_i))$$

Algorithme de construction d'un AJH

Principe

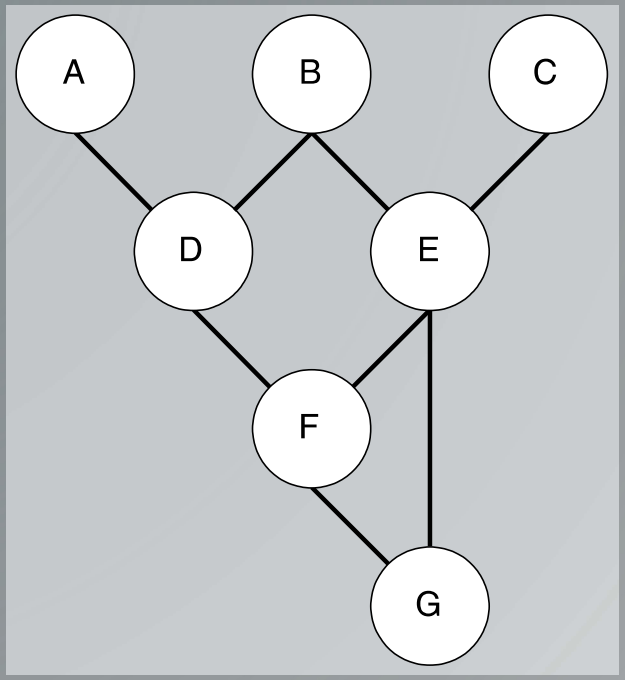
- On va construire un graphe de facteurs progressivement, croisement par croisement.
- Dès qu'un cycle apparaît, on applique une série d'opérations atomiques pour s'en débarasser en créant ou augmentant des aggrégats.

β et δ sont les deux parents, α est une variable quelconque. β^* et δ^* des branches de longueur quelconque. β^- est une branche non-vide.

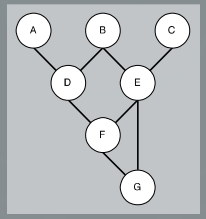


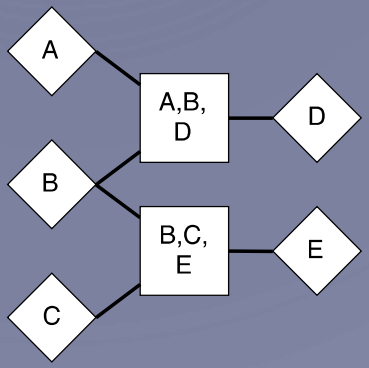
- Lorsqu'il n'y a plus de cycle, on peut continuer avec le croisement suivant.
- L'aggrégation peut engendrer des liens facteur—facteur. Il est trivial de recréer les interfaces à la fin.
- Lorsqu'on a ajouté tous les facteurs, on reprend à l'intérieur des aggrégats, en incorporant d'abord les interfaces entrantes.

Un exemple de construction

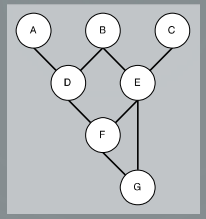


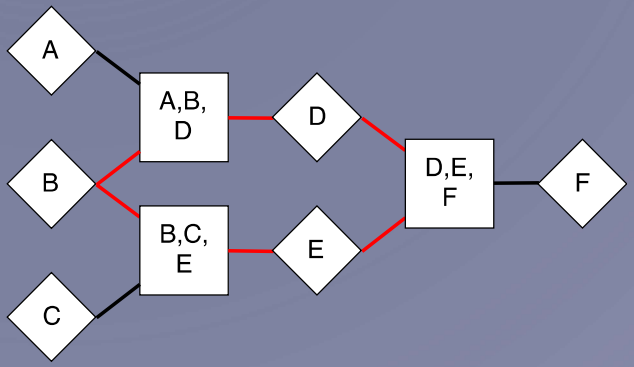
Un exemple de construction



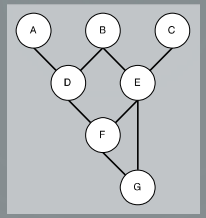


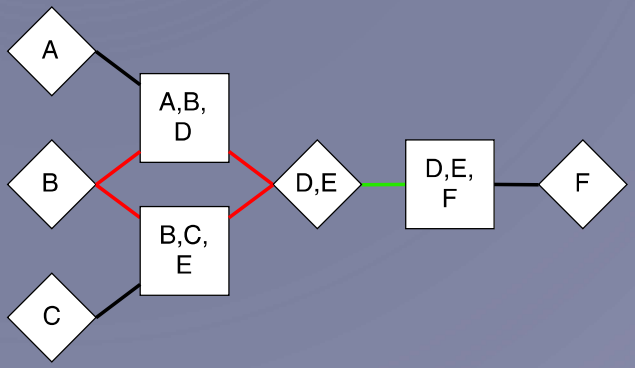
Un exemple de construction



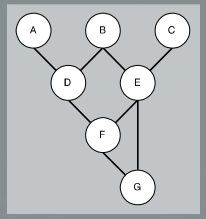


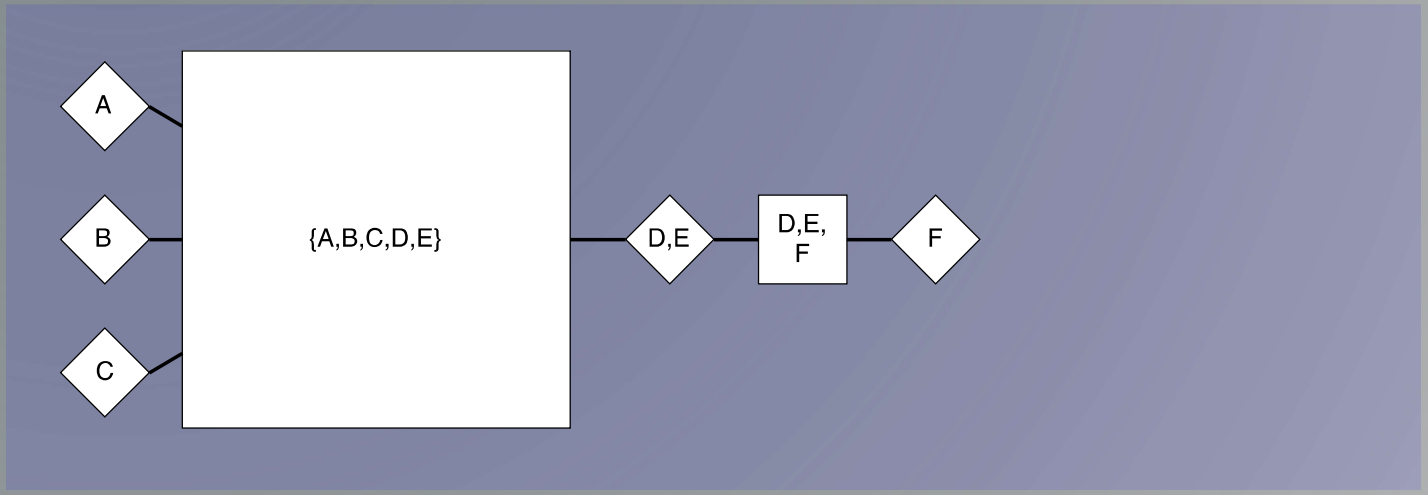
Un exemple de construction



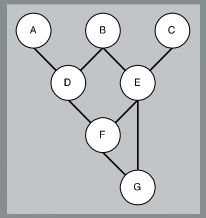


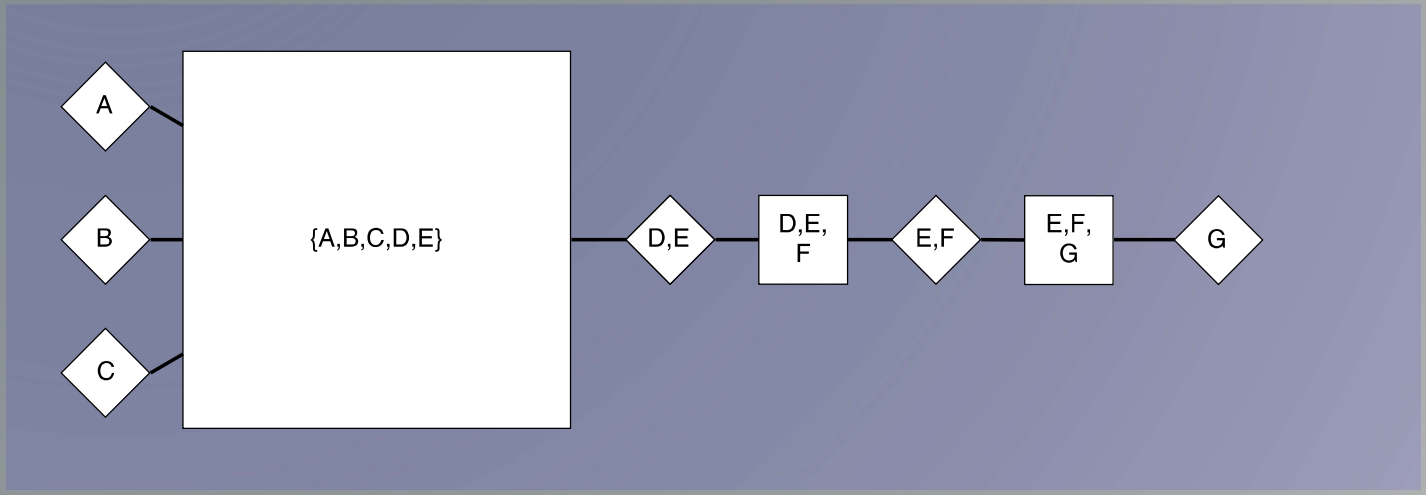
Un exemple de construction



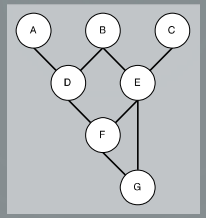


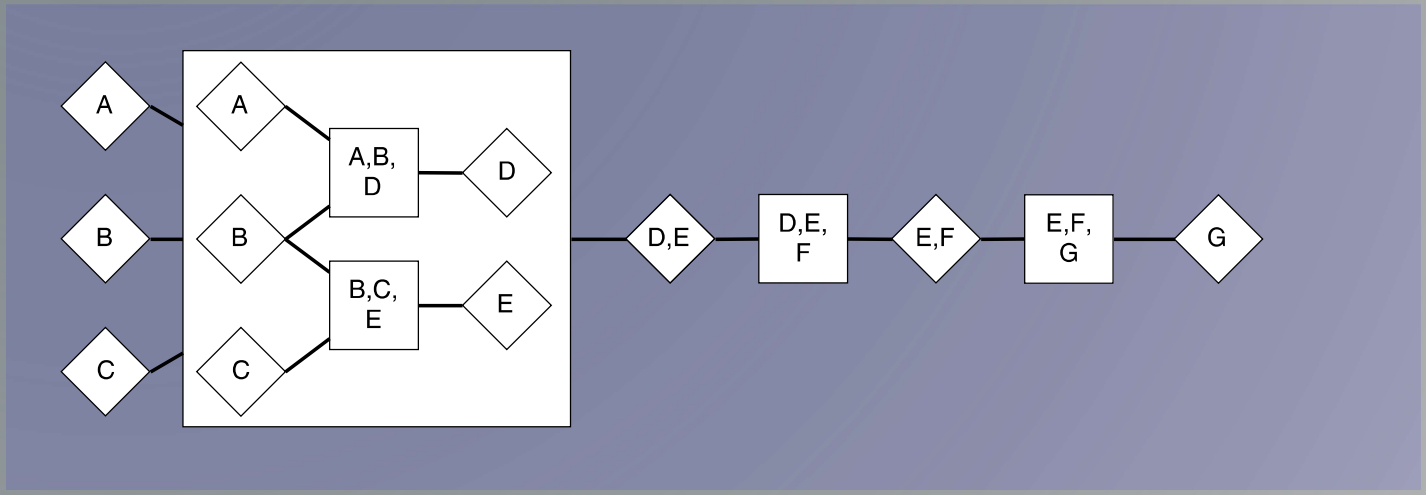
Un exemple de construction





Un exemple de construction





Inférence exacte

Une fois l'AJH construit,

⇒ Toute sous-partie a une structure d'arbre (ou de forêt.)
⇒ Les observations sont traitées comme des messages entrants.
⇒ Un simple *forward-backward* permet de calculer l'état du réseau.

Garantie limitée

- La garantie du calcul exact sans itération ne veut pas dire moins d'opérations.
- Chaque sous-graphe aggrégé devra être calculé (en profondeur) autant de fois qu'il a de voisins.

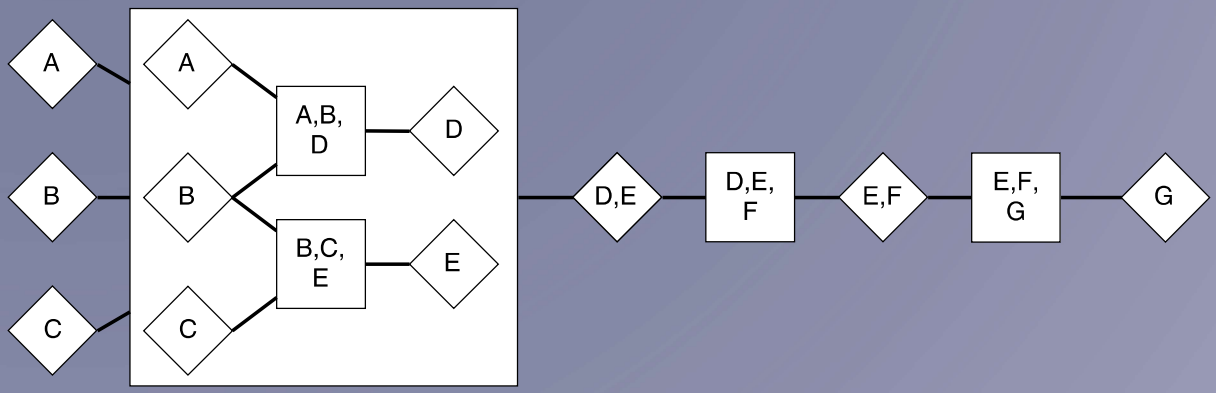
Garantie quand même

- La compacité de la représentation rend ces calculs moins coûteux.
- L'ensemble des calculs repose sur une seule opération de produit–projection.
  ⇒ L'utilisation de tables *sparse* ouvre la porte à un algorithme compact et rapide qui sort du cadre de cette présentation.
  (Je n'ai pas assez de place dans la marge)

En pratique

Messages et observations

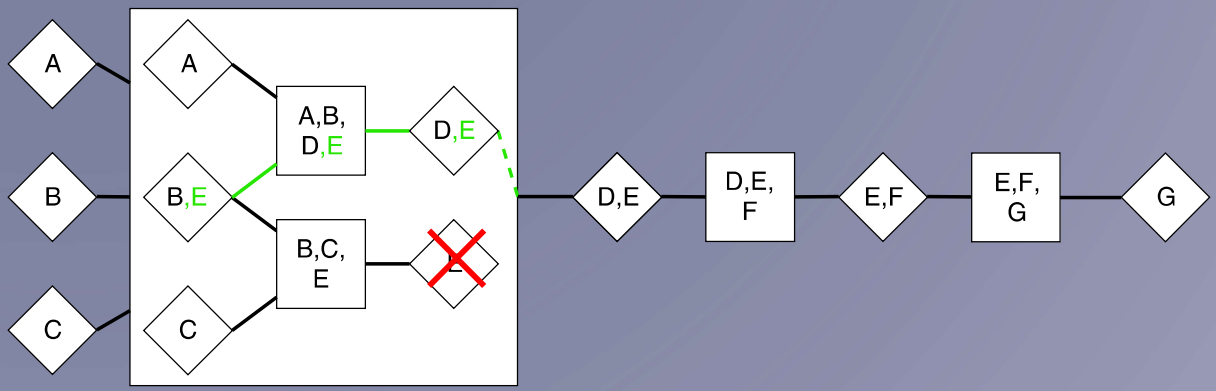
- Dans un sous-graphe donné, une observation est indiscernable d'un message entrant.
- Il est nécessaire d'ancrer chaque message entrant sur une interface donnée.
- La croyance sur une variable est extraite du facteur qui la "produit".
- Le calcul d'un message sortant revient à calculer l'état de toutes les branches entre les messages entrants ancrés et les facteurs des variables à extraire.



En pratique

Messages et observations

- Dans un sous-graphe donné, une observation est indiscernable d'un message entrant.
- Il est nécessaire d'ancrer chaque message entrant sur une interface donnée.
- La croyance sur une variable est extraite du facteur qui la "produit".
- Le calcul d'un message sortant revient à calculer l'état de toutes les branches entre les messages entrants ancrés et les facteurs des variables à extraire.



Conclusion

Nous avons implémenté cette méthode et le produit–projection optimisé dans notre analyseur de QTL Spell-QTL

https://forgemia.inra.fr/QTL/spell-qtl

Nous avons traité des données jusqu'à 2200 individus avec 8 ancêtres et 2 allèles sur 6 générations