Sélection de TAGSNP par réseau de contraintes pondérées

D. Allouche , S. Degivry, M. Sanchez, T. Schiex

Equipe SaAB : unité de Biométrie intelligence Artificielle

Centre INRA de Toulouse

plan

- Introduction du cadre méthodologique
- Problématique tagSNP
- Le modèle et les méthodes de recherche
- expérimentations

Cadre formel

Modèle Graphique **(X,D,C)**:

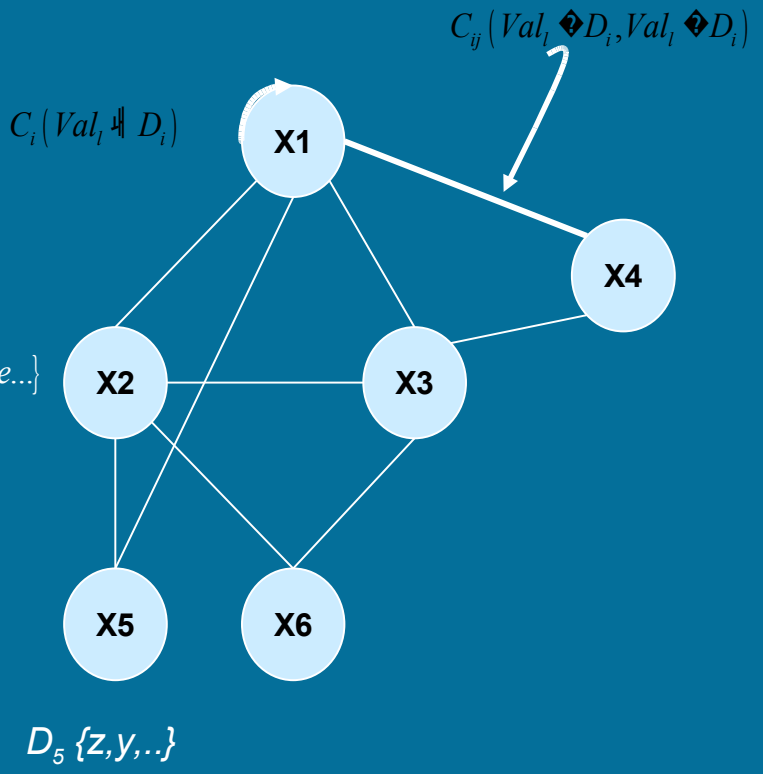
X = {Xi,…Xn} variables

D = {D1, … Dn} domaines de valeurs

C = {F1,…,Fr} contraintes=fonctions de coût

- Optimisation: trouver une affectation totale tel que:
  - $MIN \Sigma_J F_J$

(**Le problème est NP difficile)**



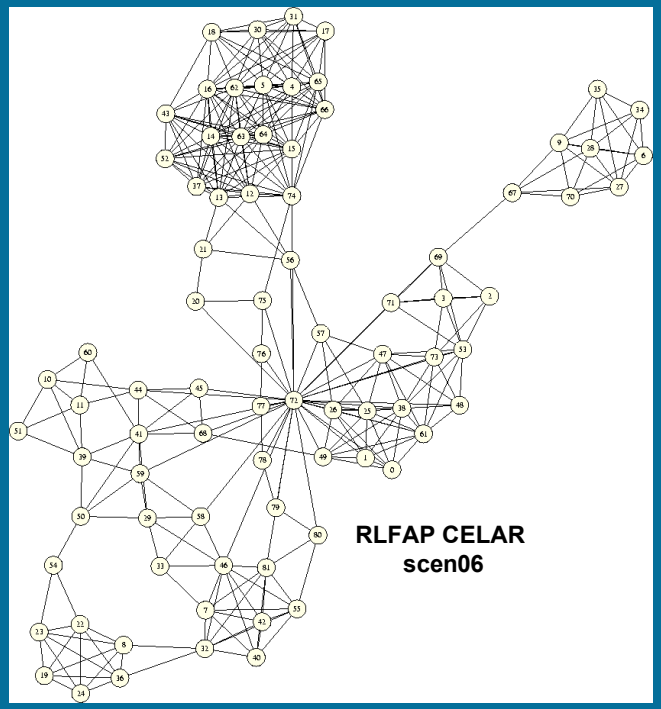
X_2	X_5	cout
a	z	1
a	y	8
..	..	∈ [0,k]
e	y	50

OPTIMISATION VIA TOULBAR2

DFBB, BTD , BTD-RDS

Un champ d'applications …large

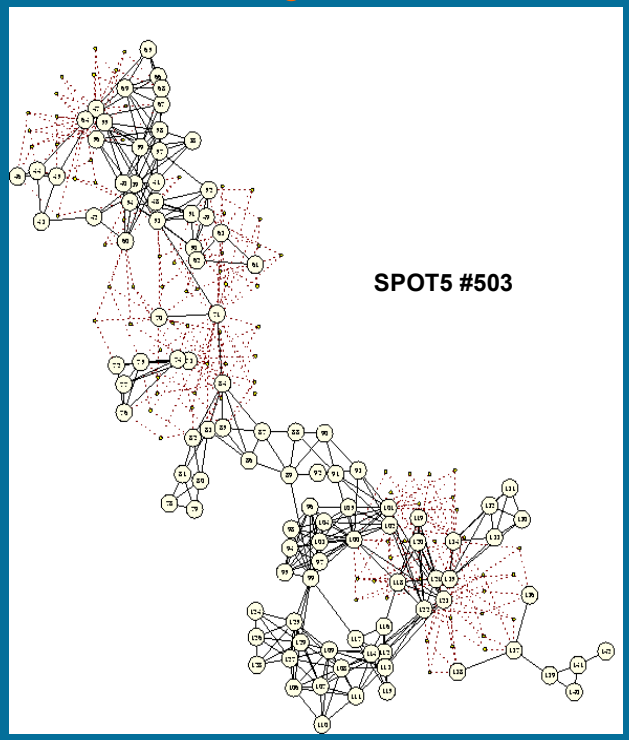
TELECOM: Radio Link Frequency Allocation Problem



Allocation de fréquences de connexion radio de sorte a minimiser les interférences

➔Minimiser contrainte binaires soft.

SPATIAL: Earth Observation Satellite Management Problem



Sélections d'images parmi un ensemble candidat en respectant les contraintes physique de la caméra.

➔ Contrainte binaire, contrainte ternaire dure.

-Maximiser la somme de poids associés aux images sélectionnées.

➔ Contraintes unaires soft

plan

- Introduction du cadre méthodologique
- **Problématique tagSNP**
- Le modèle et les méthodes de recherche
- Expérimentations

Contexte bio: Snp what is it ?

- SNP = Single Nucleotide Polymorphism

- SNP = mutation ponctuelle entre individus d'une même espèce.
  - Exemple : chez l'humain 10 M de SNP sur 6 Milliards de bases (www.hapmap.org)

- Remarques : dans une espèce, le nombre de SNP est fonction:
  - de la taille du génome
  - De la localisation (zone chromosomique, intron/exon)
  - de la diversité génétique de l'espèce

– Les SNP sont majoritairement hérités des parents

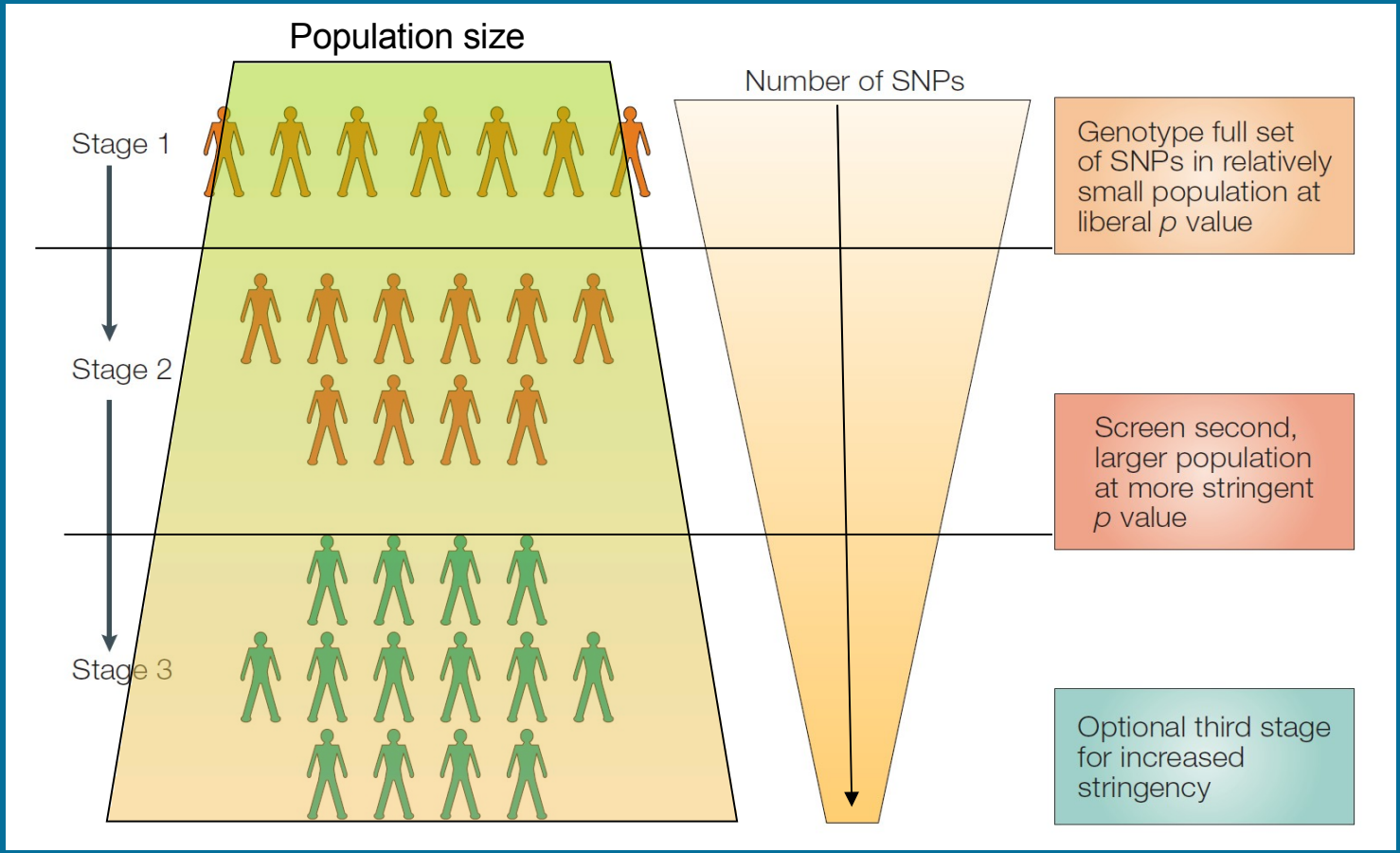
Contexte Bio: Des SNP et alors ?

- SNP sont impliqués dans les maladies ou caractères multifactoriels
  - Susceptibilité à des maladies
  - Efficacité de médicaments
  - Réponse quantitative de caractères d'intérêt agronomique

Étude d'association par déséquilibre de liaison est une approche de recherche de causalité entre : SNP et phénotype d'intérêt

➔ Cout expérimental encore prohibitif! ➔ sélection de tagSNP

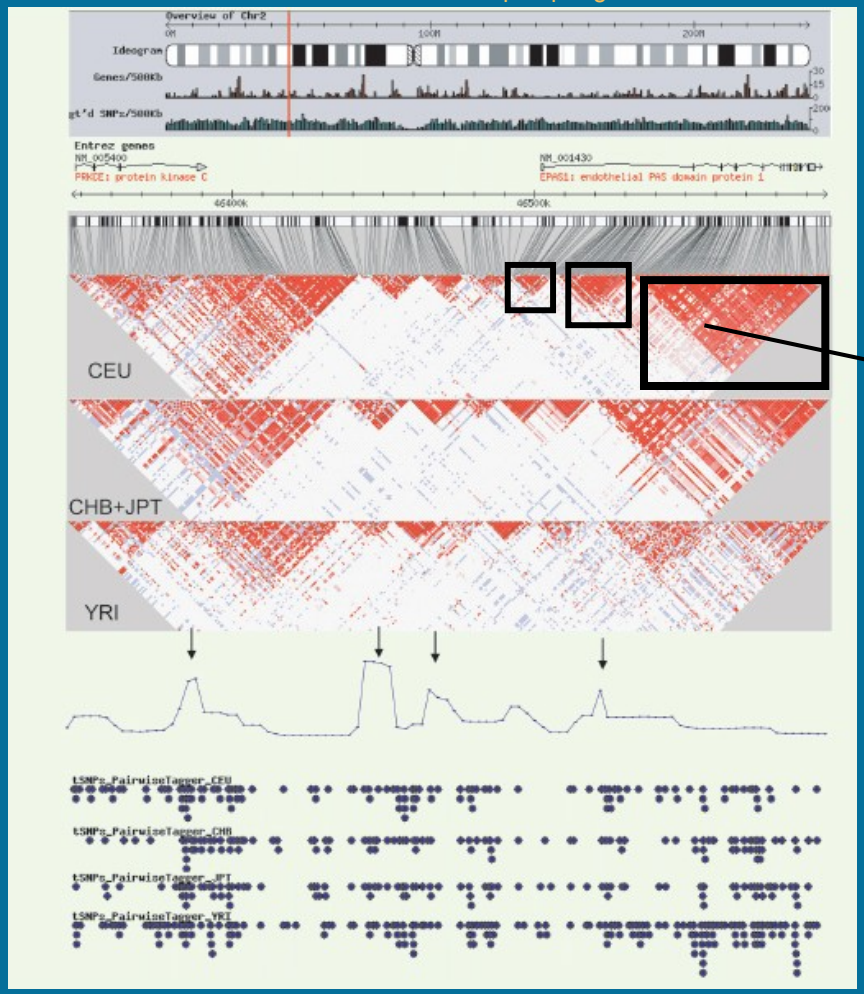
Stratégie pour accroître l'efficacité de l'étude d'association



« Hirschhorn NAT. REV. GEN. vol 6 p95 2005 »

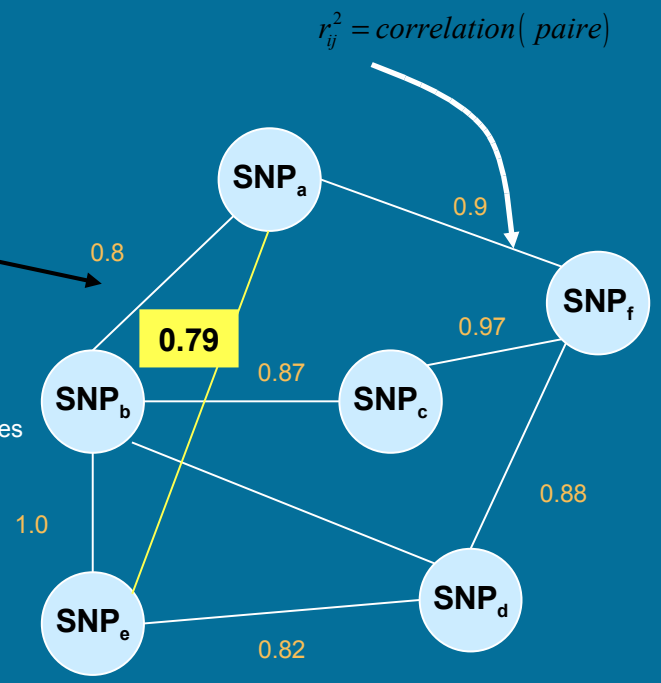
Carte haplotypique

Carte HAPMAP: wwww.hapmap.org



Représentation Graphique :

Plusieurs composantes connexes



- Filtrage des arêtes $r^2_{ij} \geq r^2_{cut}$

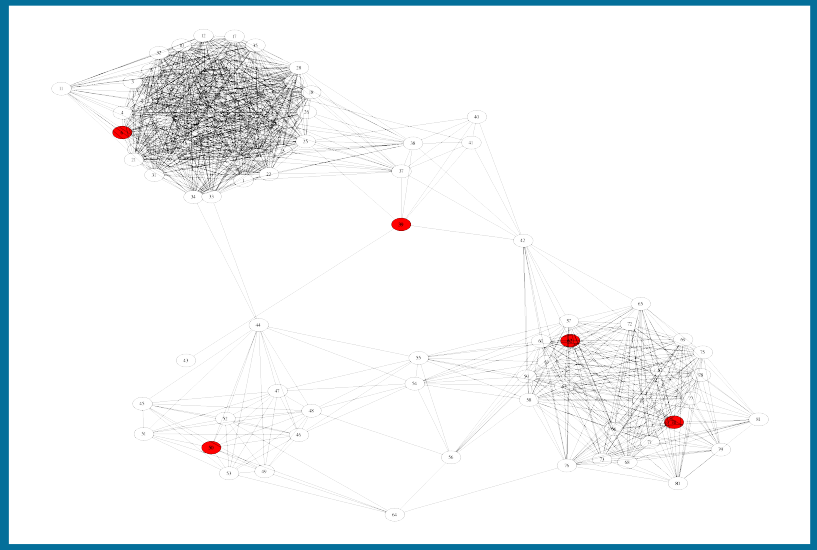
($r^2_{cut} = 0.8$)

CEU: CEPH (Utah residents with ancestry from northern and western Europe)

CHB: Han Chinese in Beijing, China+**JPT:** Japanese in Tokyo, Japan

YRI: Yoruba in Ibadan, Nigeria

Définition du problème « tagSNP »



Sélectionner le long d'un chromosome un sous ensemble de marqueurs SNP de taille minimale

de sorte que les marqueurs sélectionnés (tagSNP) soient les plus représentatifs de l'information génétique de l'ensemble.

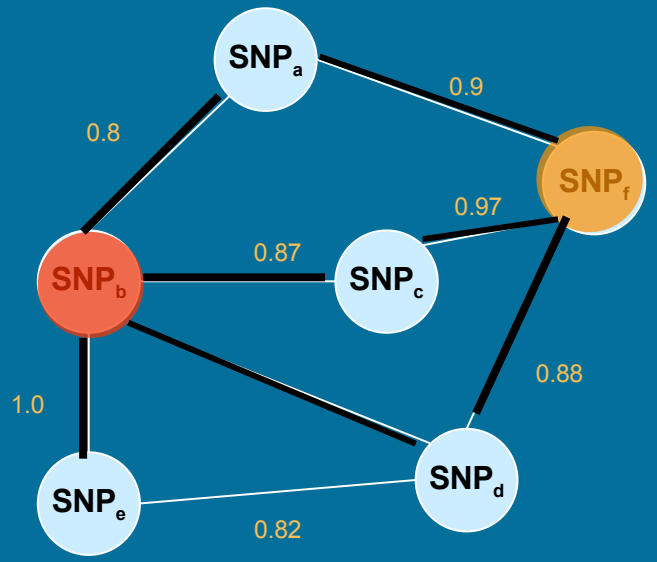
➔Problème de "set covering" avec des critères additionnels.
(problème NP-difficile)

Sélection de Tagsnp ⇔ set covering

Le set covering est un problème classique en informatique et théorie de la complexité.

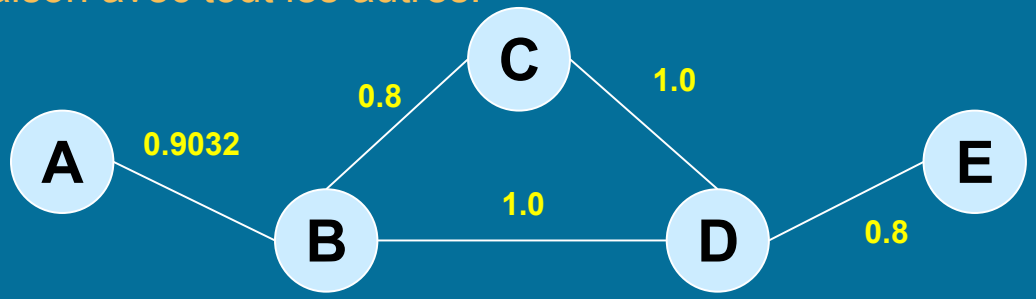
- Soit une ensemble composé de plusieurs familles.
  (avec éventuellement des éléments commun).

- L'objectif est de sélectionner le nombre minimum de famille afin de recouvrir tous les éléments compris dans l'ensemble total.



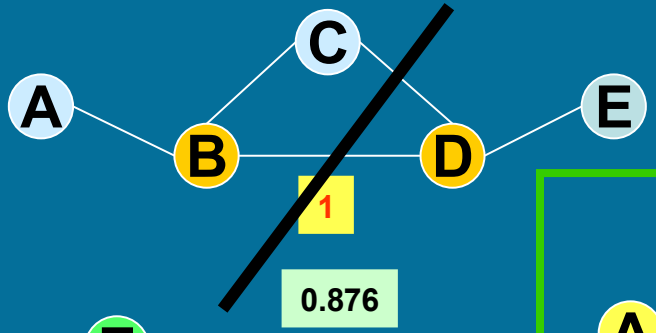
Des solutions optimales multiples:

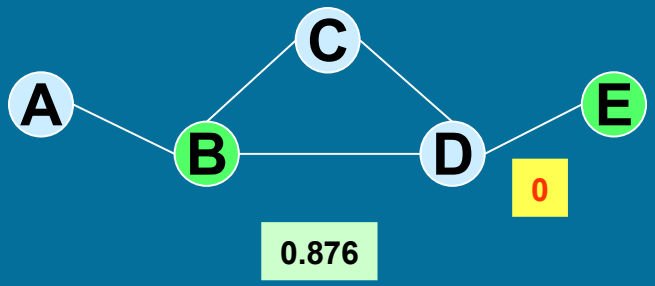
Recherche de Tagsnp = recherche d'un sous ensemble de taille minimal de snp en déséquilibre de liaison avec tout les autres.

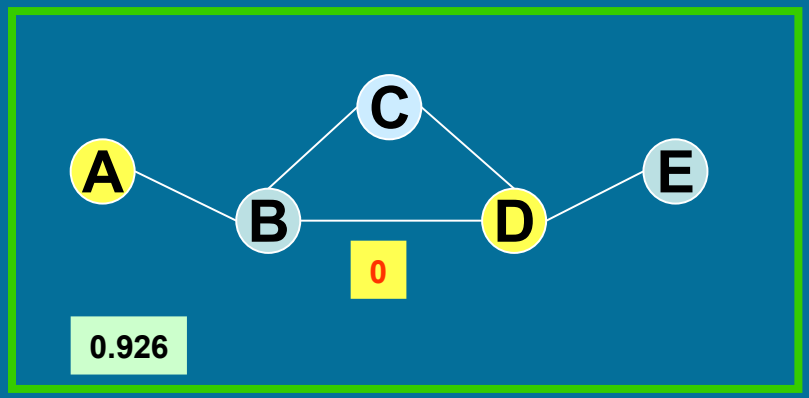


DISPERSION:
r^2 moyen entre tag

REPRESENTATIVITE:
r^2 moyen entre tag et non tag





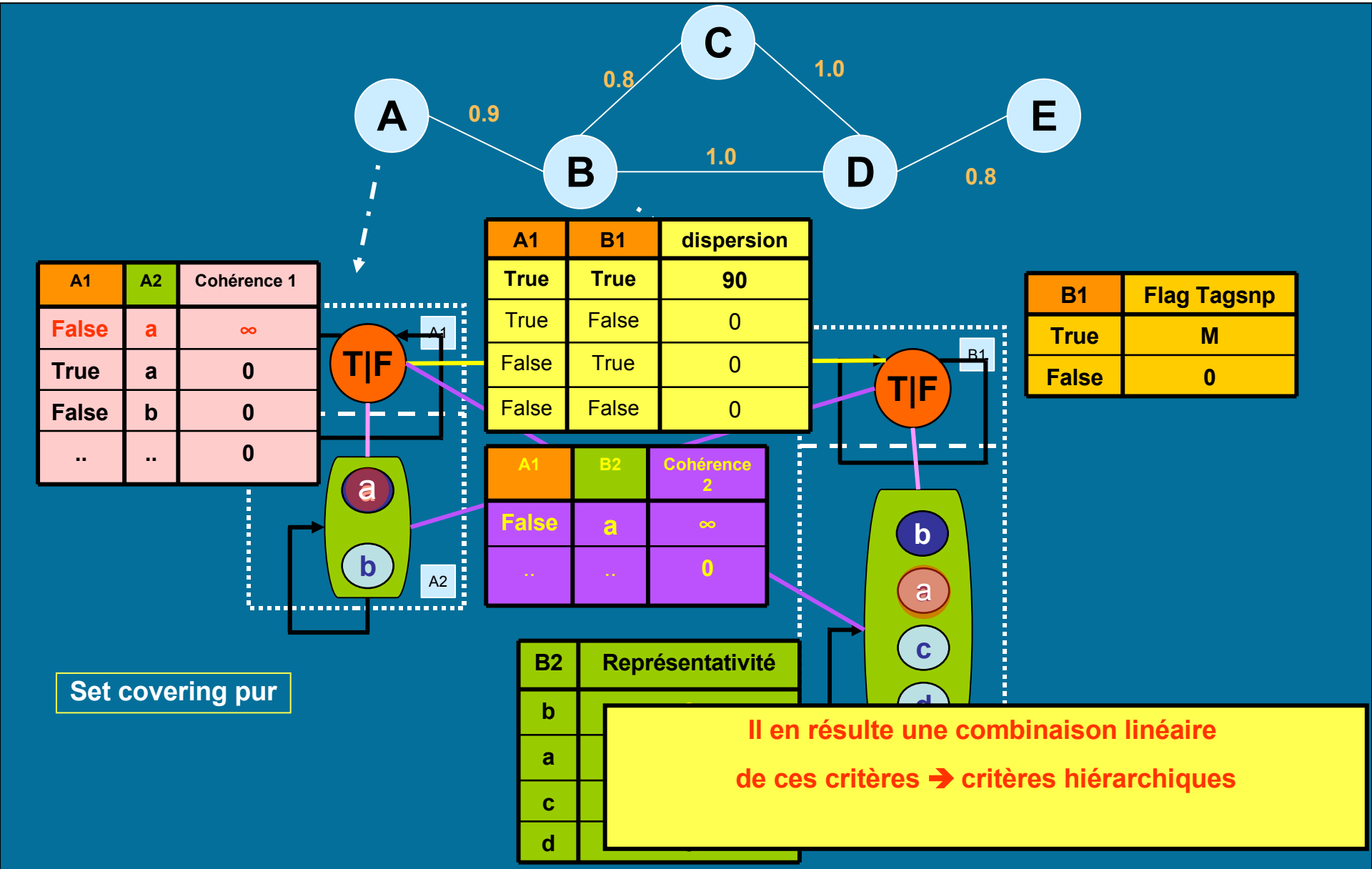


Solution = {(B,E) | (B,D) | (A,D) }

plan

- Introduction du cadre méthodologique
- **Problématique tagSNP**
- Le modèle et les méthodes de recherches
- Expérimentations

Le reseau de fonction de cout



A1	A2	Cohérence 1
False	a	∞
True	a	0
False	b	0
..	..	0

A1	B1	dispersion
True	True	90
True	False	0
False	True	0
False	False	0

B1	Flag Tagsnp
True	M
False	0

A1	B2	Cohérence 2
False	a	∞
..	..	0

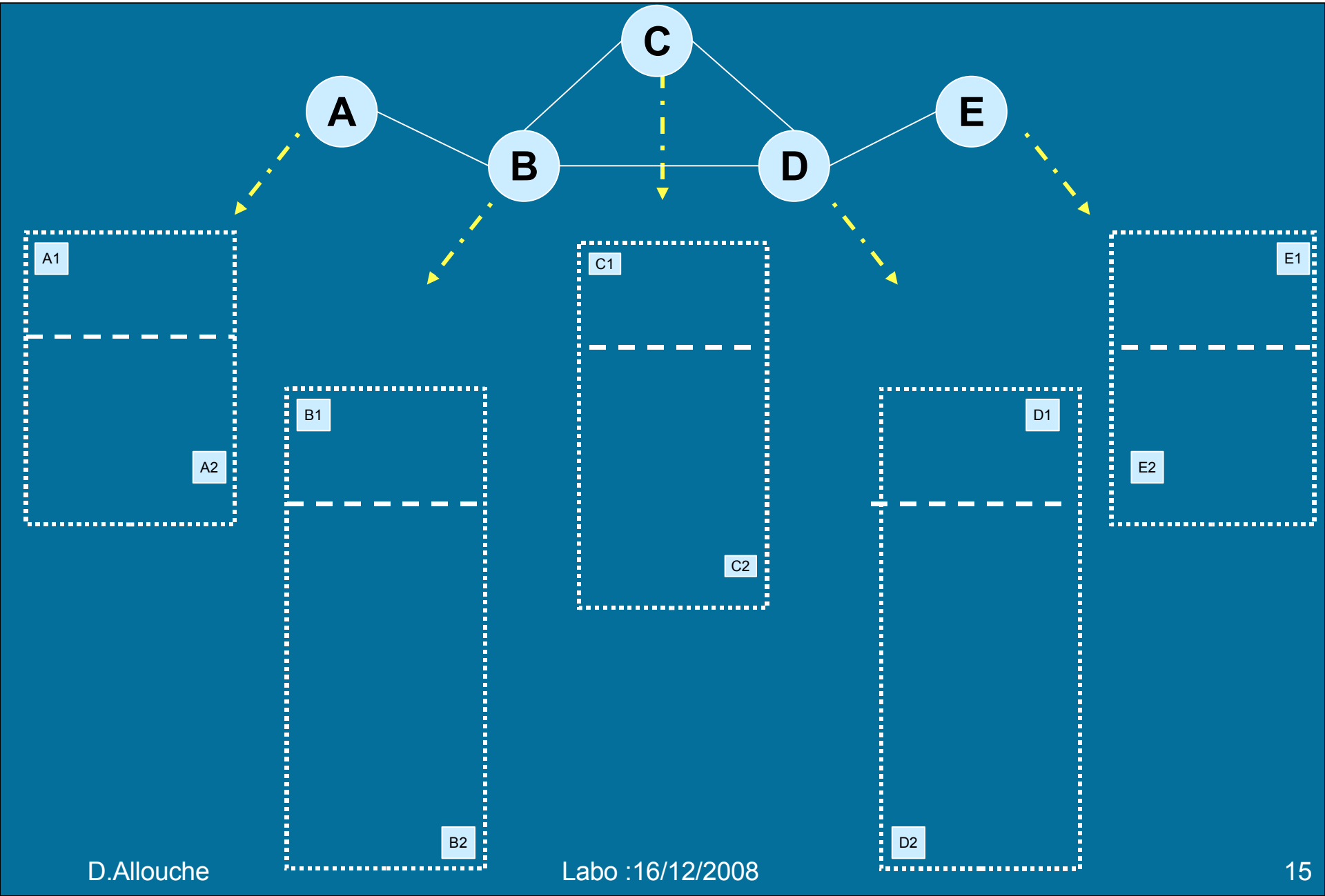
B2	Représentativité
b	
a	
c	
d	

Set covering pur

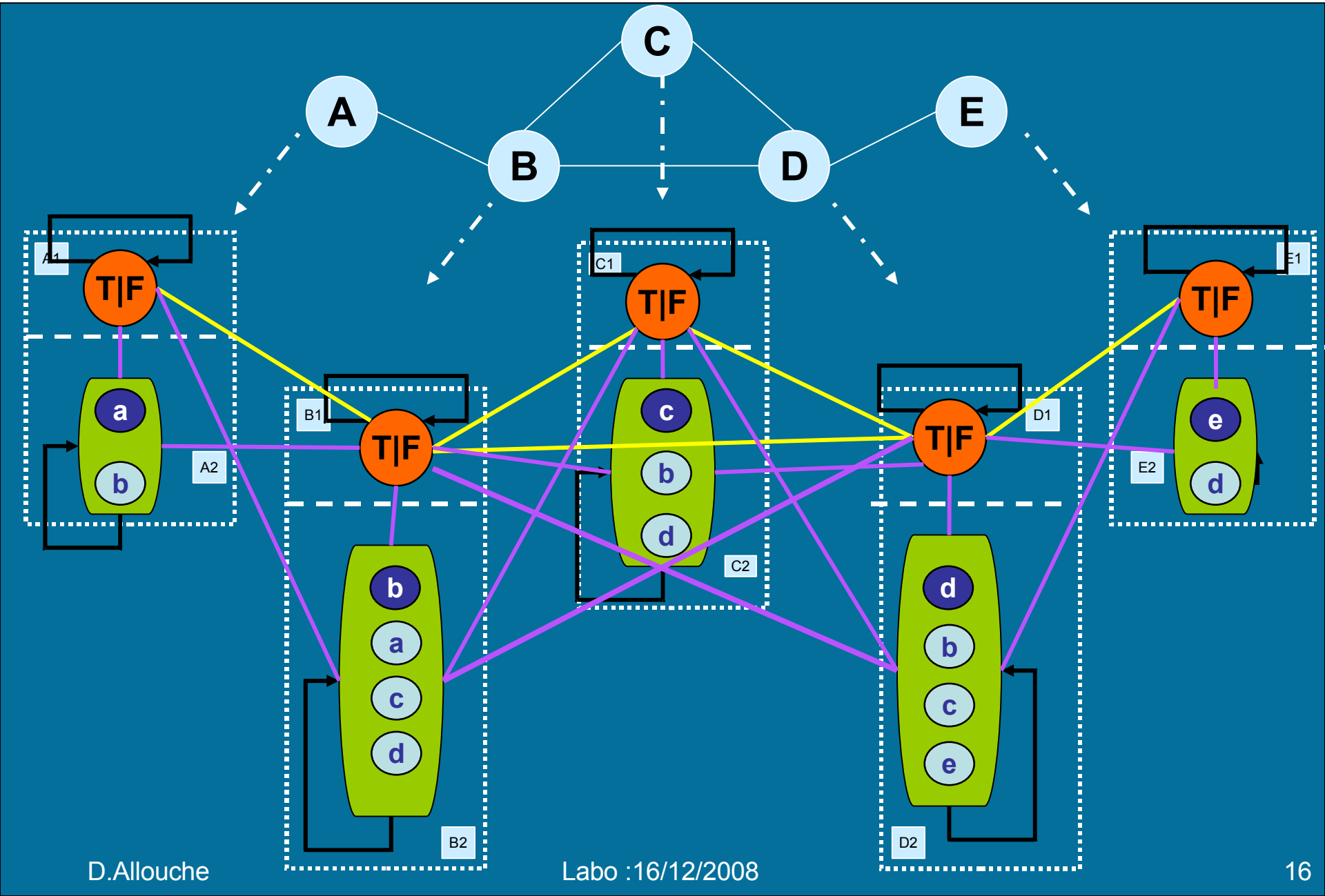
Il en résulte une combinaison linéaire

de ces critères ➔ critères hiérarchiques

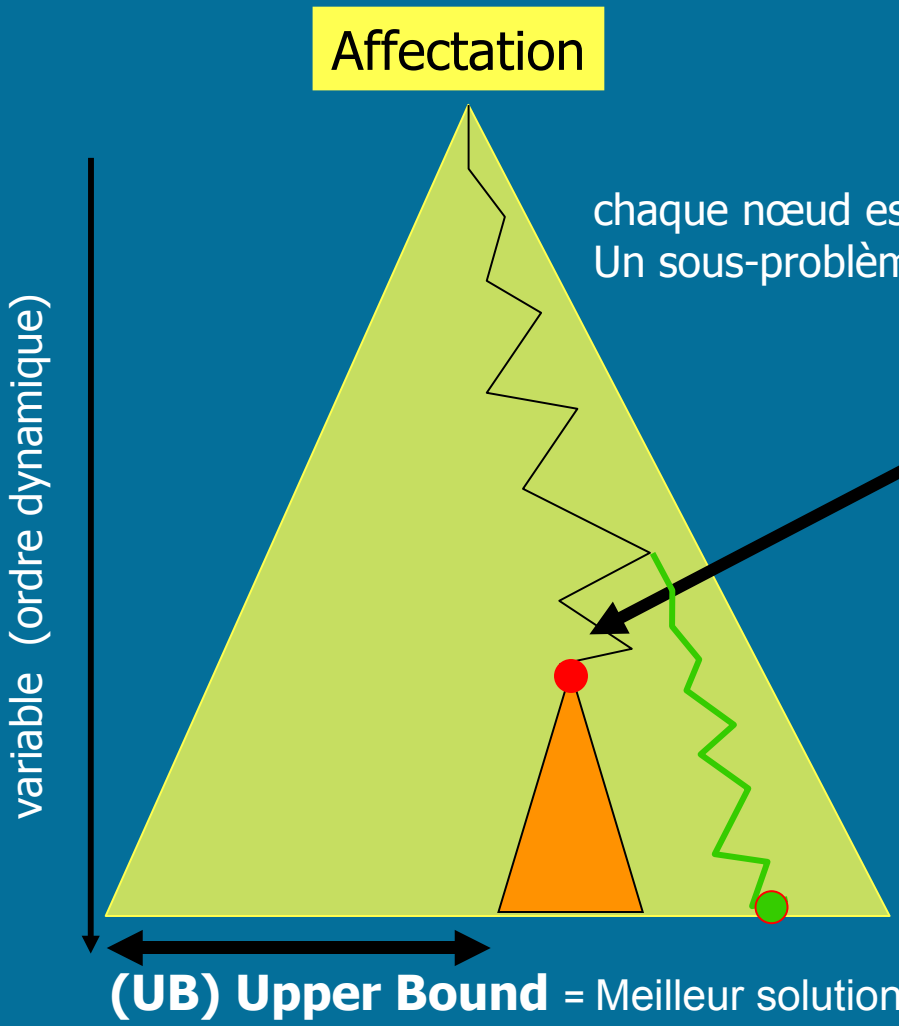
Le Modèle



Le Modèle


C
A
B
D
E
A1
T|F
a
b
A2
B1
T|F
b
a
c
d
B2
C1
T|F
c
b
d
C2
D1
T|F
d
b
c
e
D2
E1
T|F
e
d
E2

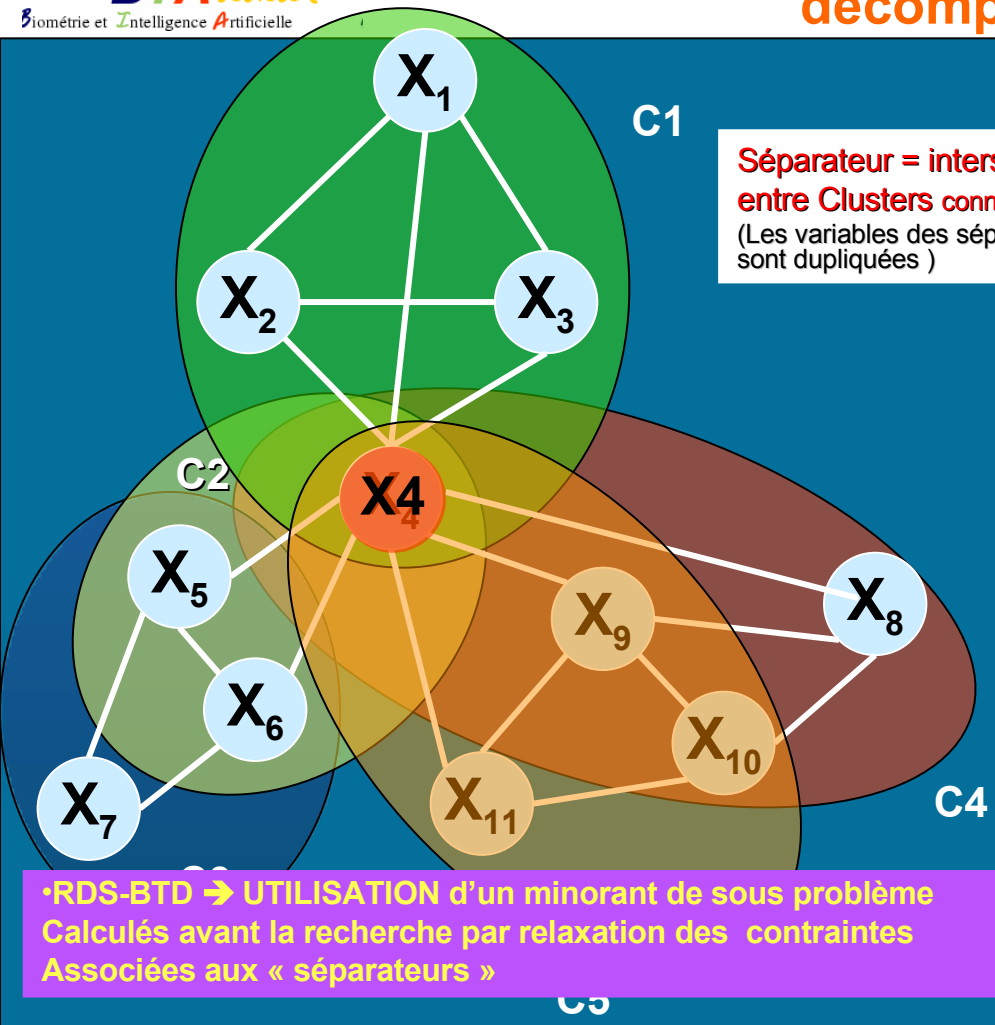
Depth-First Branch and Bound (DFBB)



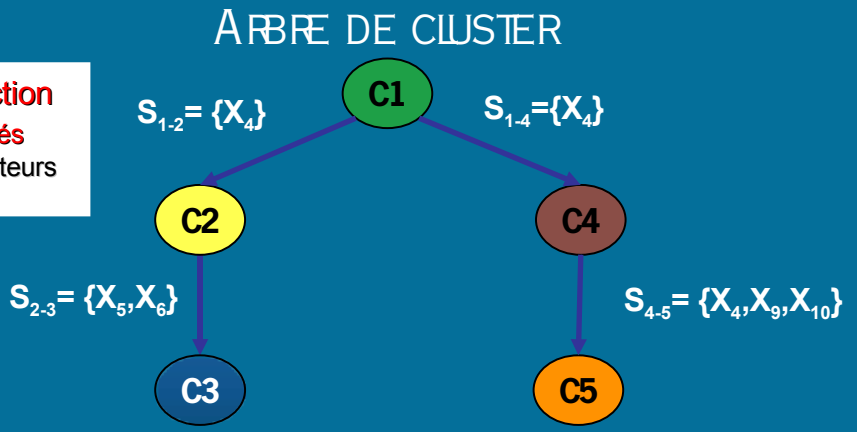
Time complexity: $O(d^n)$

Space complexity: $O(n)$

Méthodes de recherches arborescentes par décomposition



Séparateur = intersection entre Clusters connectés
(Les variables des séparateurs sont dupliquées)



BTD (**B**acktrack **T**ree **D**ecomposition)

- L'affectation des séparateurs rend les sous-problèmes indépendants
- On mémorise de l'information dans les séparateurs afin d'éviter des calculs redondant

•RDS-BTD ➔ UTILISATION d'un minorant de sous problème Calculés avant la recherche par relaxation des contraintes Associées aux « séparateurs »

A partir d'un ordre d'élimination des variables on cherche successivement les cliques de tailles maximales dans le graphe triangulé.

➔ construction d'un arbre de clusters

•**Complexité :**

Time: $O(d^w)$,

Space: $O(d^s)$

d = taille du plus grand domaine
w = taille du plus grand cluster - 1 (tree width)
S = taille du plus grand séparateur

plan

- Introduction du cadre méthodologique
- Problématique tagSNP
- Le modèle et les méthodes de recherche
- Expérimentations

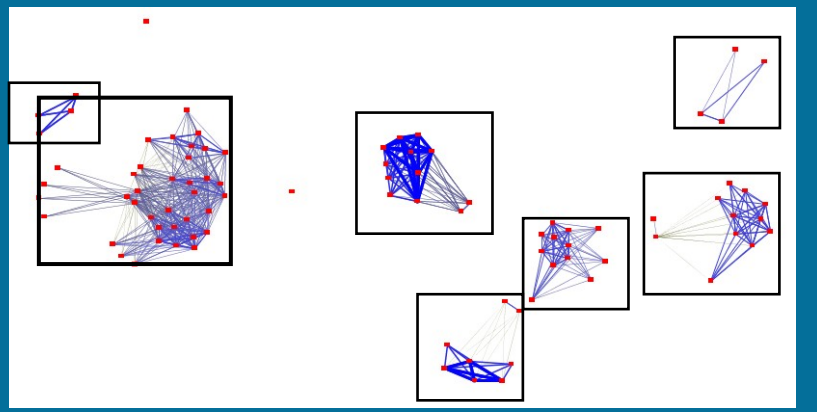
Expérimentation: Données de départ

Chromosome 1 : Hapmap phase 2

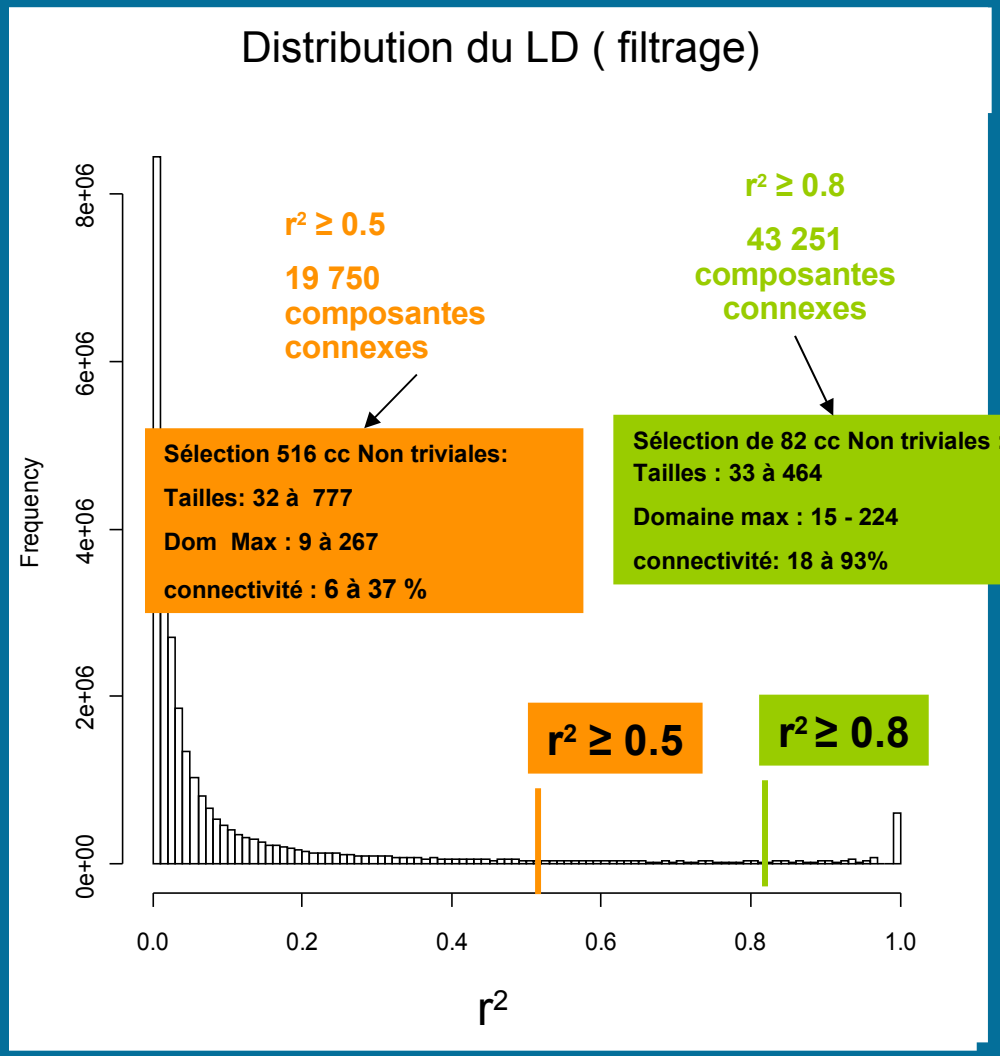
46 Sujets Européens

178 438 SNP

30 Millions d'arrêtes



Composantes connexes =instances indépendantes



Logiciel de référence

- FESTA (QIN et al. Bioinformatics Nov. 2005):
  - Approche exhaustive pour les composantes connexes de petites tailles
  - Recherche gloutonnes pour les composantes connexes de plus grandes tailles.

- « Aujourd'hui la grande majorité des logiciels de sélection de tagSNP sont basés sur des méthodes incomplètes ».

Résultats d'expérimentations

- $r^2 \geq 0.8$
  - DFBB ➔Résolution de la totalité des instances (82/82)

    Comparaison de l'approche à DFBB à « FESTA gloutonne »

    - Temps de calcul : 2h37 avec DFBB versus 3mm avec FESTA
    - Gain de la compression de -21 % par rapport FESTA (#tagsnp passe de 487 à 359)

    Comparaison DFFB à FESTA « hybride recherche exhaustive + gloutonne »

    - Gain de la compression -3 % par rapport à FESTA (#Tagsnp passe 370 à 359)
    - Temps de calcul : 2h37 avec DFBB versus 39 h 17m FESTA
      (15 fois plus rapide que FESTA)

➢ Comparaison des méthodes de recherche Toulbar2:

- ➢ (limite 2h cpus par instances)
- DFBB ➔ 100%,
- BTD ➔ 75% (62/82)
- RDS-BTD ➔79% (65/82)

BENCHMARK

- $r^2 \geq 0.5$
  - Résolution de 95% des instances (491/516)

    Gain de compression -15% par rapport à FESTA « glouton »

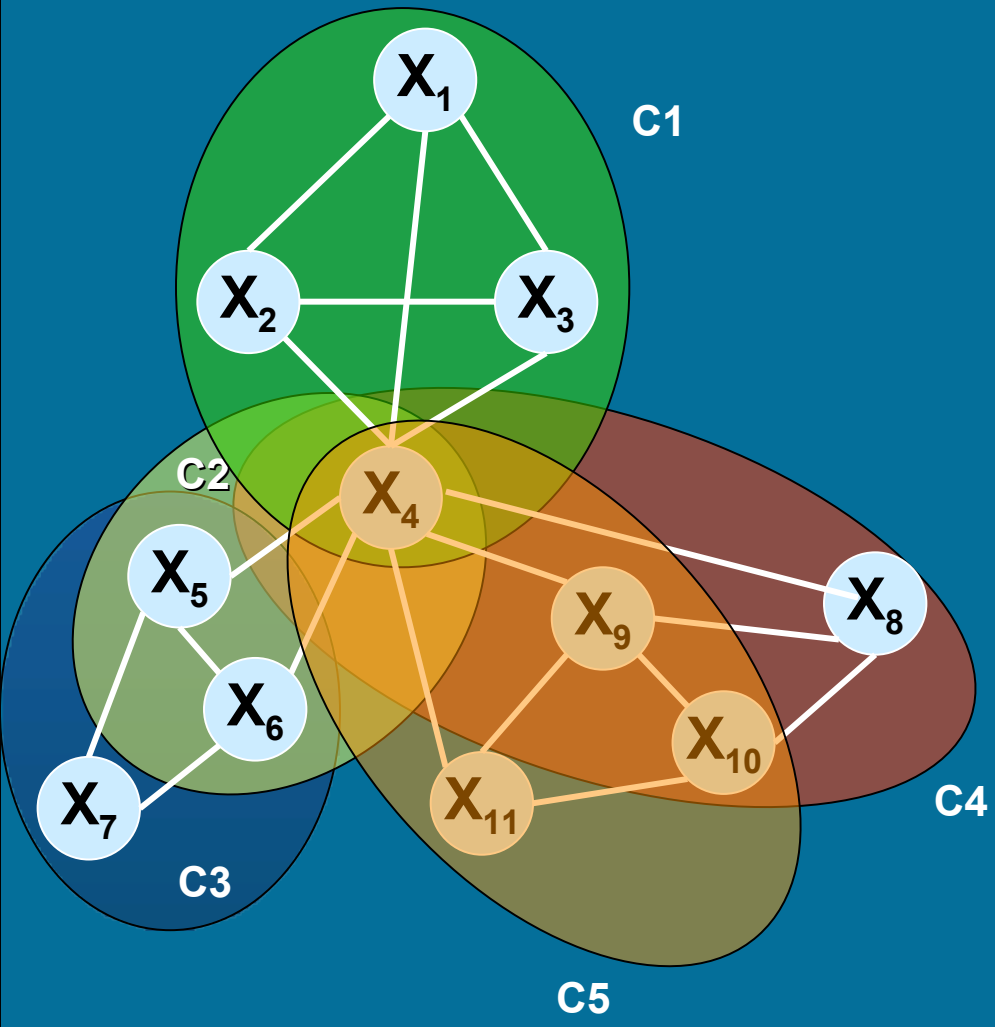
    #TAGSNP passe de 2780 versus 3274 avec FESTA

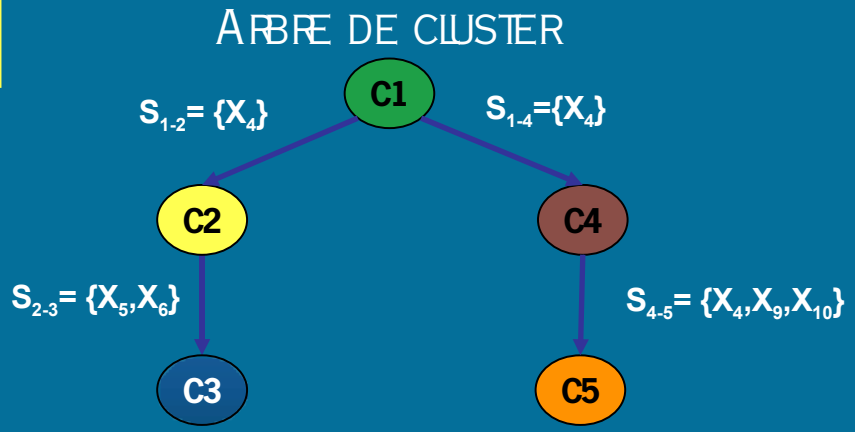
    25 instances de plus grande taille sont particulièrement difficiles
    ➔ benchmark de méthodes

  - décomposition totale des problèmes est souvent pénalisante!

Modification de la décomposition : recherche de petits séparateurs

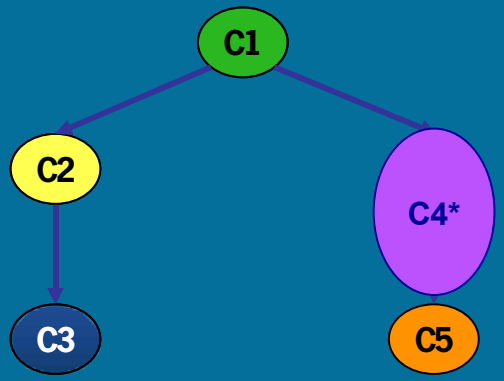
Plus de liberté dans l'ordre dynamique des variables lors de la recherche





Fusion de cluster

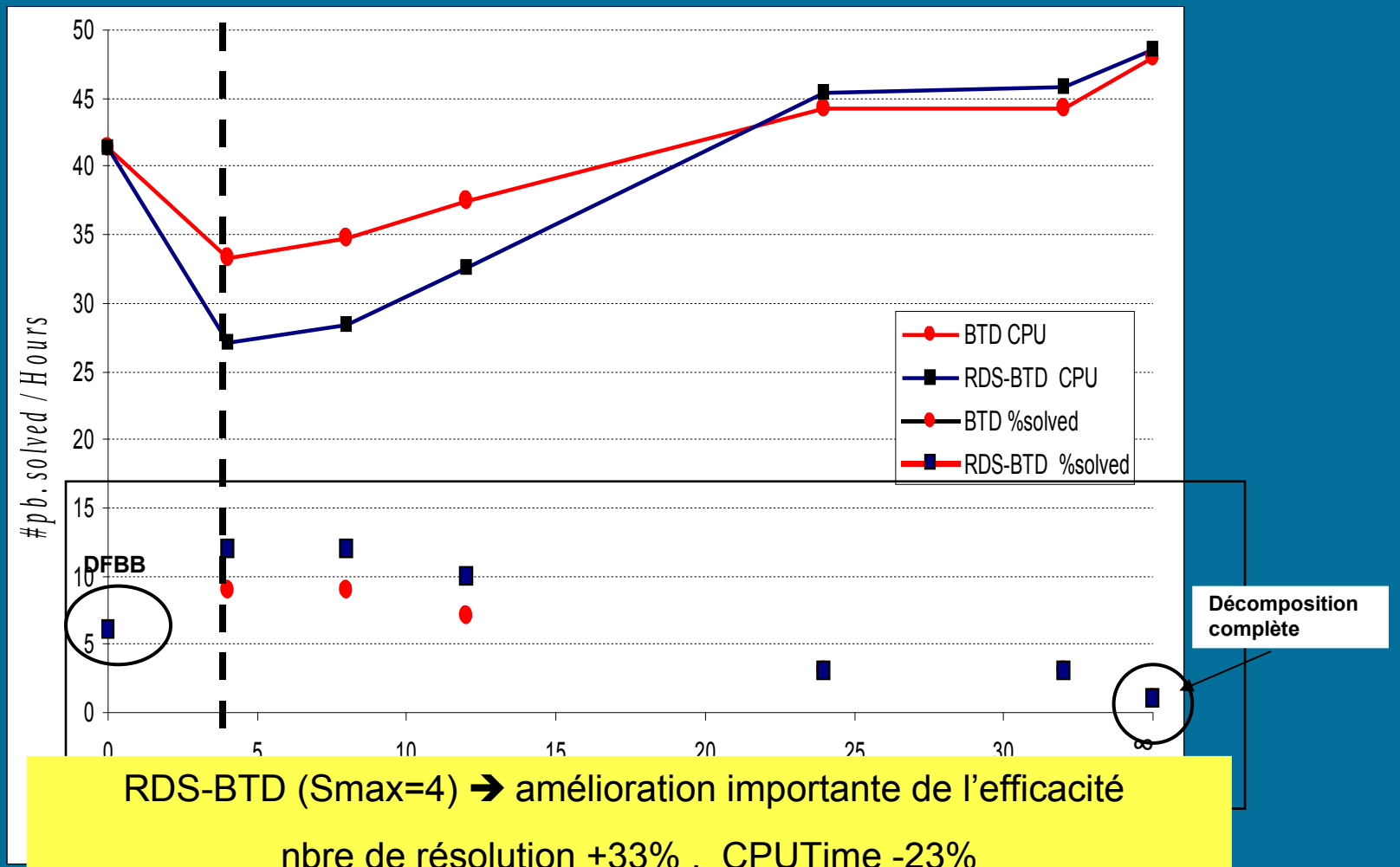
Si $\#S_{i-j} > \#S_{seuil}$



Impactes du majorant de taille de séparateur

Jeux de départ :Les 25 instances difficiles ($r^2 \geq 0.5$)

2 heures de calcul maximum



RDS-BTD (Smax=4) ➔ amélioration importante de l'efficacité

nbre de résolution +33% , CPUTime -23%

En résumé:

- Pour $r^2 \geq 0.8$: 100% des instances sont résolues (DFBB, BTD (Smax=4), RDS-BTD (Smax=4))

- Pour $r^2 \geq 0.5$:
  - 12/25 instances « difficiles » restent ouvertes
  - Intérêt pour la biologie modéré ! ➔ benchmark méthode d'optimisation

- Concernant les méthodes:
  - RDS-BTD améliore les résultats par rapport à BTD.
  - Décomposition avec petit séparateur facilite la résolution des instances

perspectives

- Extension des méthodes de découpage de graphes Toulbar:
  - Intégration de méthode de découpage permettant exhiber les petits séparateurs avec une prise en compte de la structure et les cout.
    - Algo découpage de graphe type metis, hmetis
    - Algo fusion/fission de graphe (Brichot&al)
    - Algo de découpage par relaxation lagrangienne
    - ..?

THE END …..

TOULBAR2:

http://mulcyber.toulouse.inra.fr/gf/project/toulbar2

…

Quelques références:

DFBB **➔*AND/OR search*** **(Marinescu & Dechter, 2005)**

BTD (Simon de Givry, Thomas Schiex, and Gérard Verfaillie) In /Proc. of AAAI-06/, Boston,2006

RDS **(Verfaillie *et al*, 1996)** ***Pseudo-Tree RDS*** **(Larrosa *et al*, 2002)**