Antoine-Lorquin Aymeric, équipe Dyliss, IRISA/INRIA

Orthocis

Une base de données pour l'étude des facteurs de transcriptions

Identification in silico de gènes ciblés par un facteur de transcription donné

Présentation du projet 1/3

- **Contexte : ANR Fatinteger (2012-2015)** *INRA – IRISA/INRIA*
  Florence Gondret
  - Étude du métabolisme des lipides chez les Eucaryotes supérieurs
  - Régulation des gènes chez le porc et le poulet

- **Projet Orthocis**
  - UMR PEGASE, INRA Rennes : Sandrine Lagarrigue, Frédéric Lecerf
  - DYLISS, INRIA/IRISA Rennes : Aymeric Antoine-Lorquin, Catherine Belleannée, François Moreews, Jacques Nicolas, Anne Siegel
  - But : Rechercher des gènes ciblés par un facteur de transcriptions (FT) donné
  - Moyen : Analyser tous les gènes orthologues parmi plusieurs espèces modèles et non-modèles pour identifier des sites de fixation conservés du facteur de transcription (TFBS)

Présentation du projet 2/3

- **Point de départ :** Un motif de fixation d'un FT d'intérêt
- **Objectif :**
  - Limiter l'avalanche de faux positifs lors des *pattern matching* de TFBS plein génome

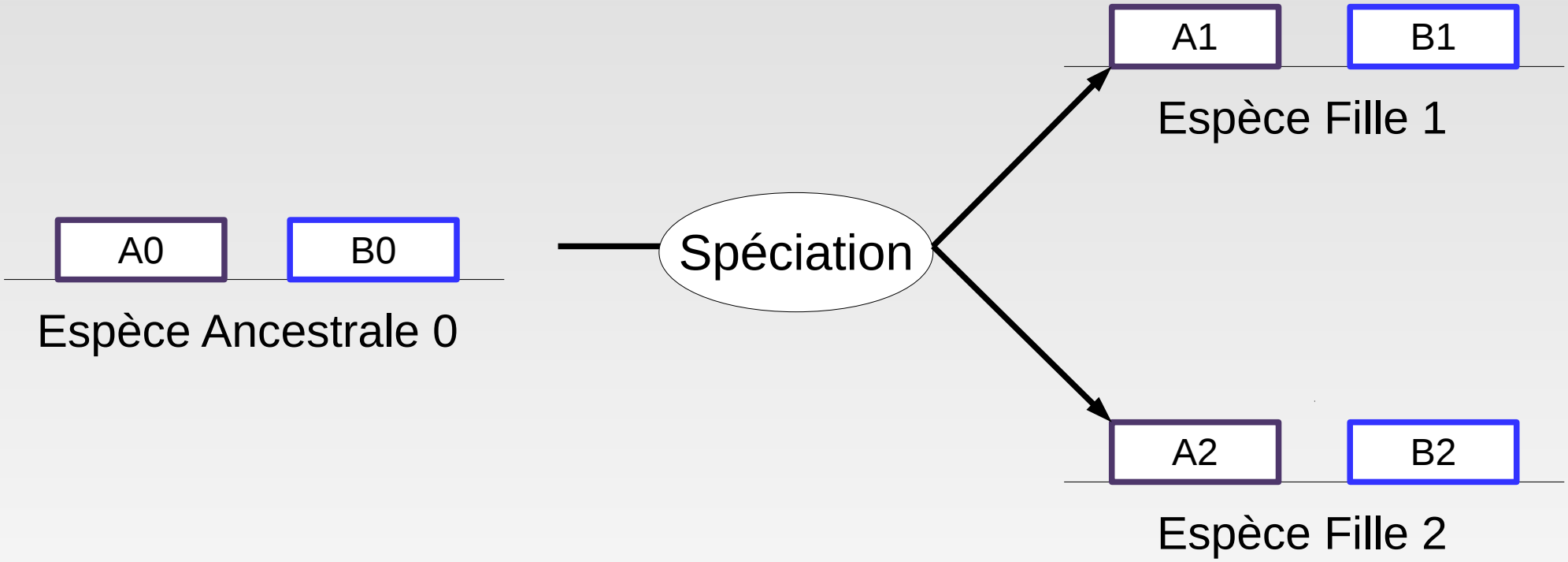
Présentation du projet 3/3

- **Point de départ :** Un motif de fixation d'un FT d'intérêt
- **Objectif** :
  - Limiter l'avalanche de faux positifs lors des *pattern matching* de TFBS plein génome
- **Principe** :
  - Exploiter l'hypothèse de conservation des TFBS au travers de l'évolution
  - **Ne valider un TFBS que s'il est présent et bien conservé dans d'autres gènes orthologues**

Création d'une base dédiée :
Orthocis

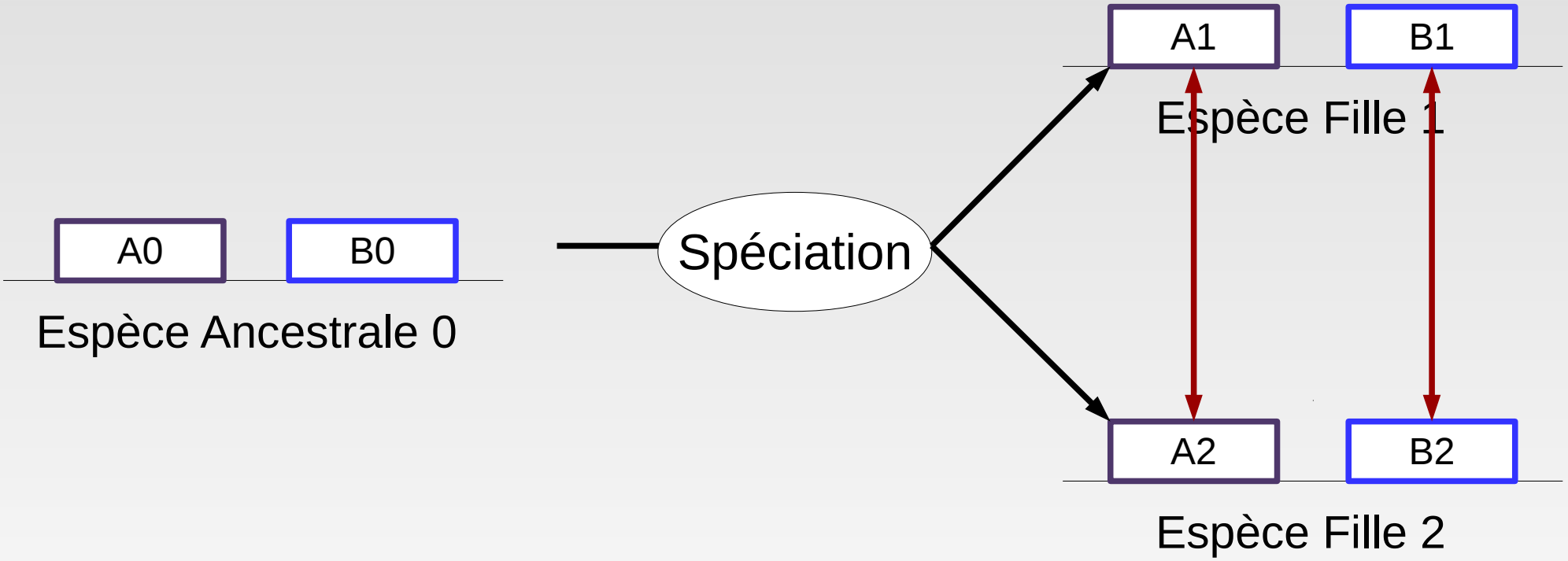
Orthologie et paralogie 1/7

- ***Orthologie (one2one)***



Orthologie et paralogie 2/7

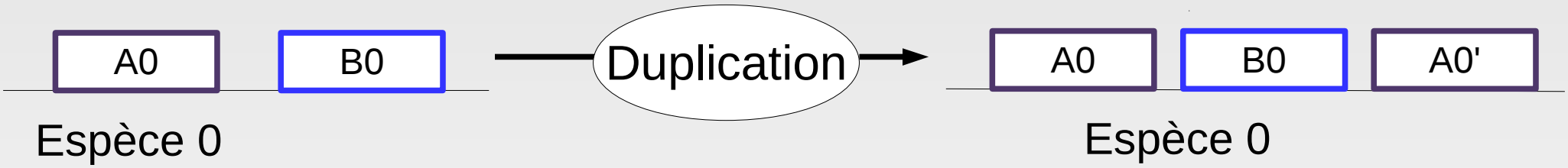
- ***Orthologie (one2one)***



Orthologie (one2one)

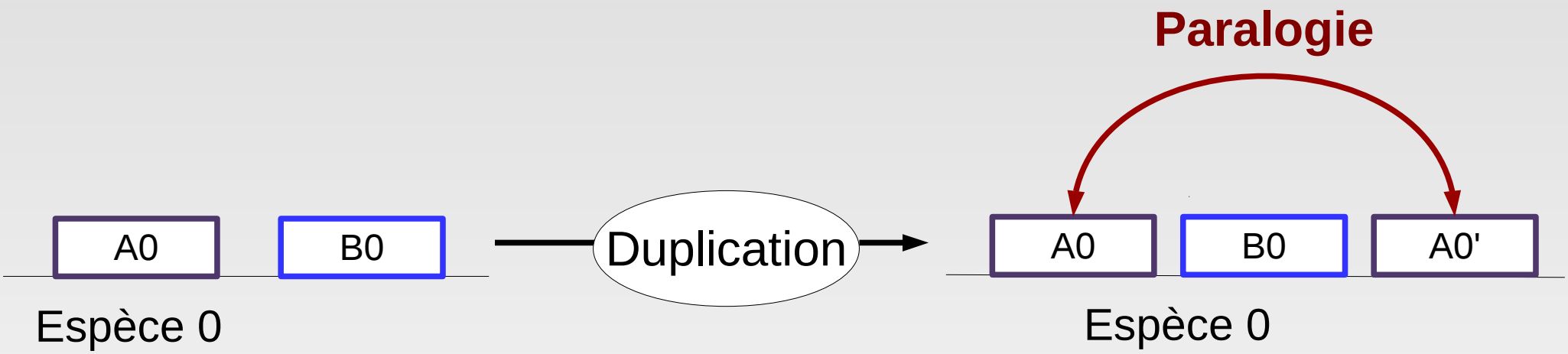
Orthologie et paralogie 3/7

- ***Paralogie***



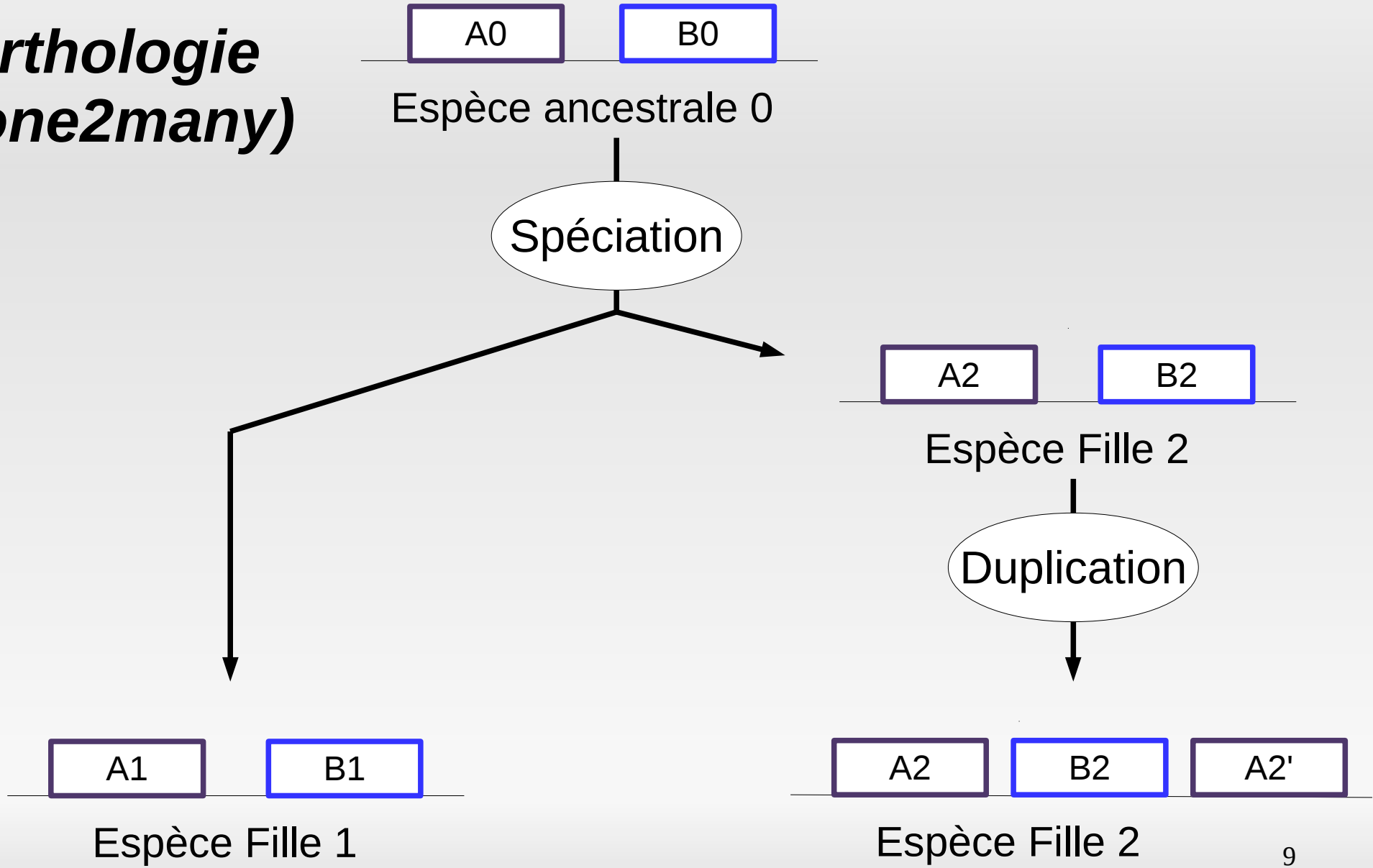
Orthologie et paralogie 4/7

- ***Paralogie***



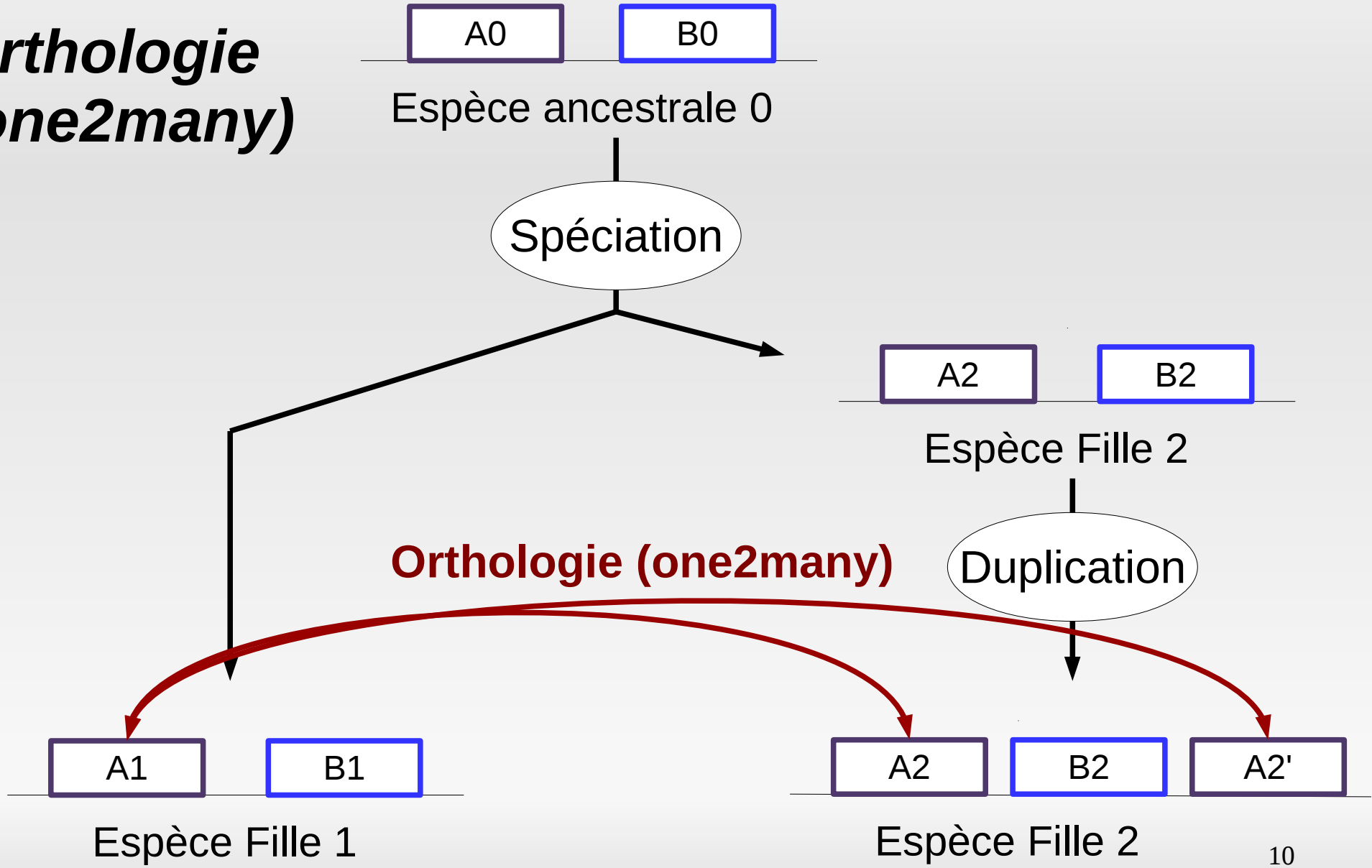
Orthologie et paralogie

- ***Orthologie (one2many)***



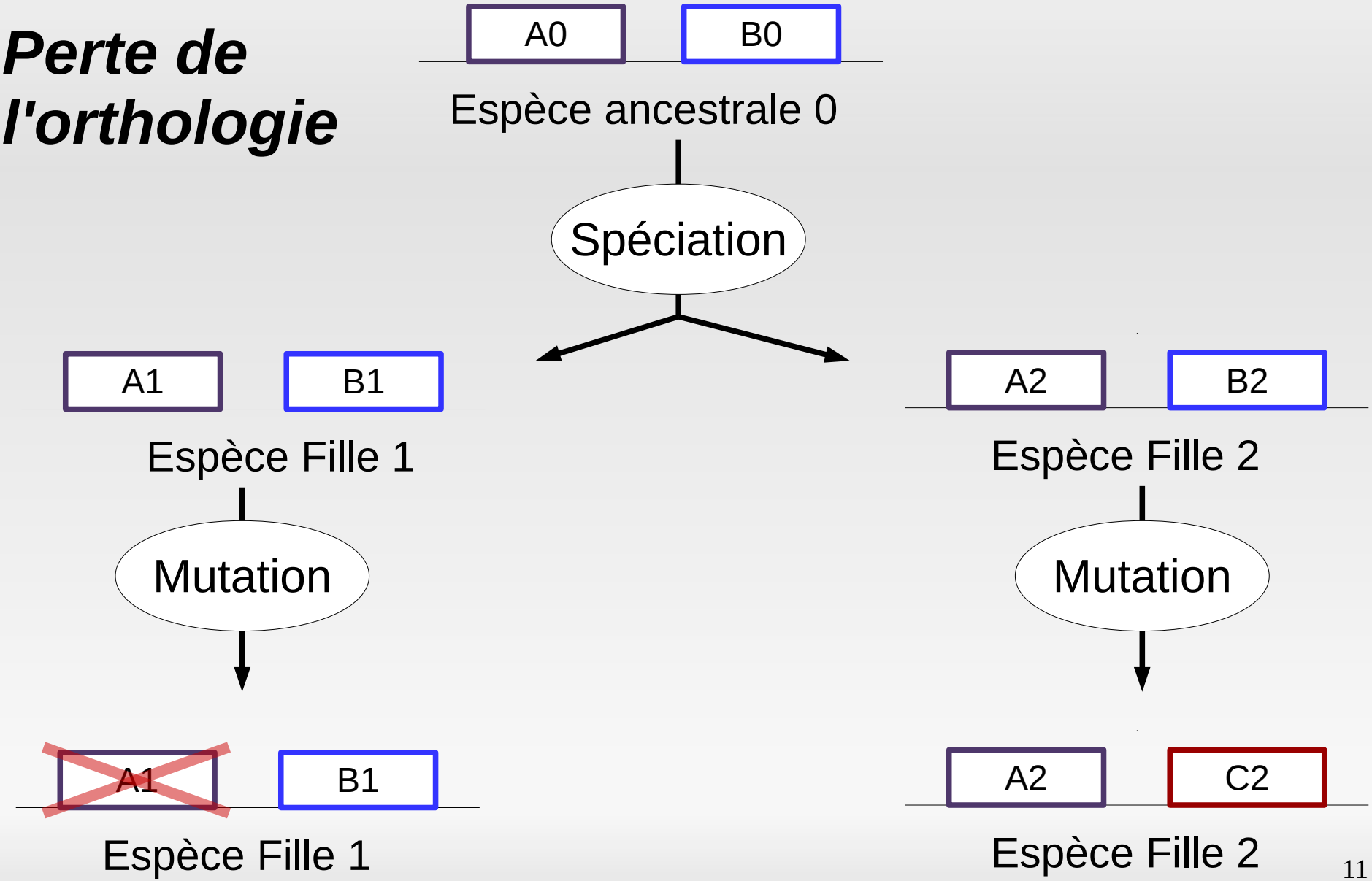
Orthologie et paralogie 6/7

- ***Orthologie (one2many)***



Orthologie et paralogie 7/7

- ***Perte de l'orthologie***



Orthocis, une base de données 1/2

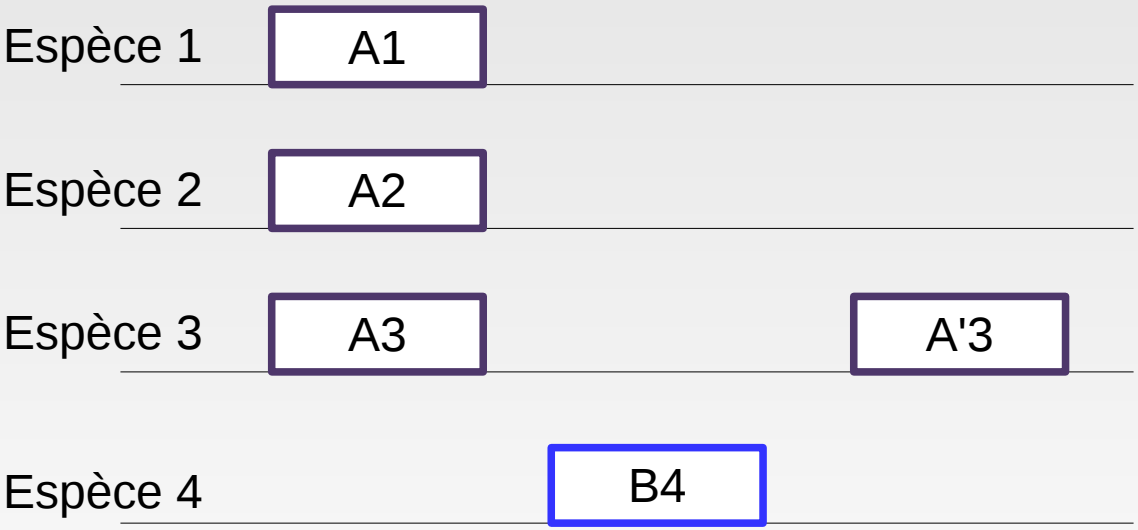
- **Informations stockées dans la base de données** :
  - La liste de tous les gènes eucaryotes d'Ensembl présentant au moins un lien d'orthologie one2one vers une autre espèce
  - Les zones upstreams de 10 000 nt de ces gènes

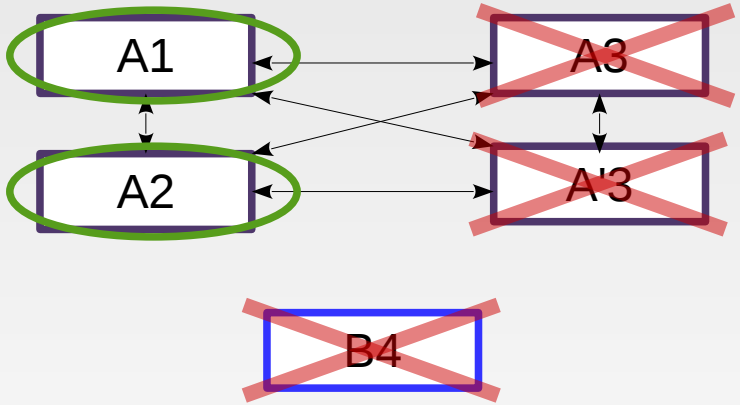
    1 063 411 gènes répartis entre 65 espèces
  - Des motifs de TFBS
  - Pour un motif donné, tous les hits obtenus sur tous les upstreams

    ex : FT LXR chez l'humain = 86 060 hits
- **Outils de pattern matching utilisés**
  - RSAT, pour la recherche de motifs sous forme matricielle
  - D'autres outils peuvent facilement être intégrés à Orthocis

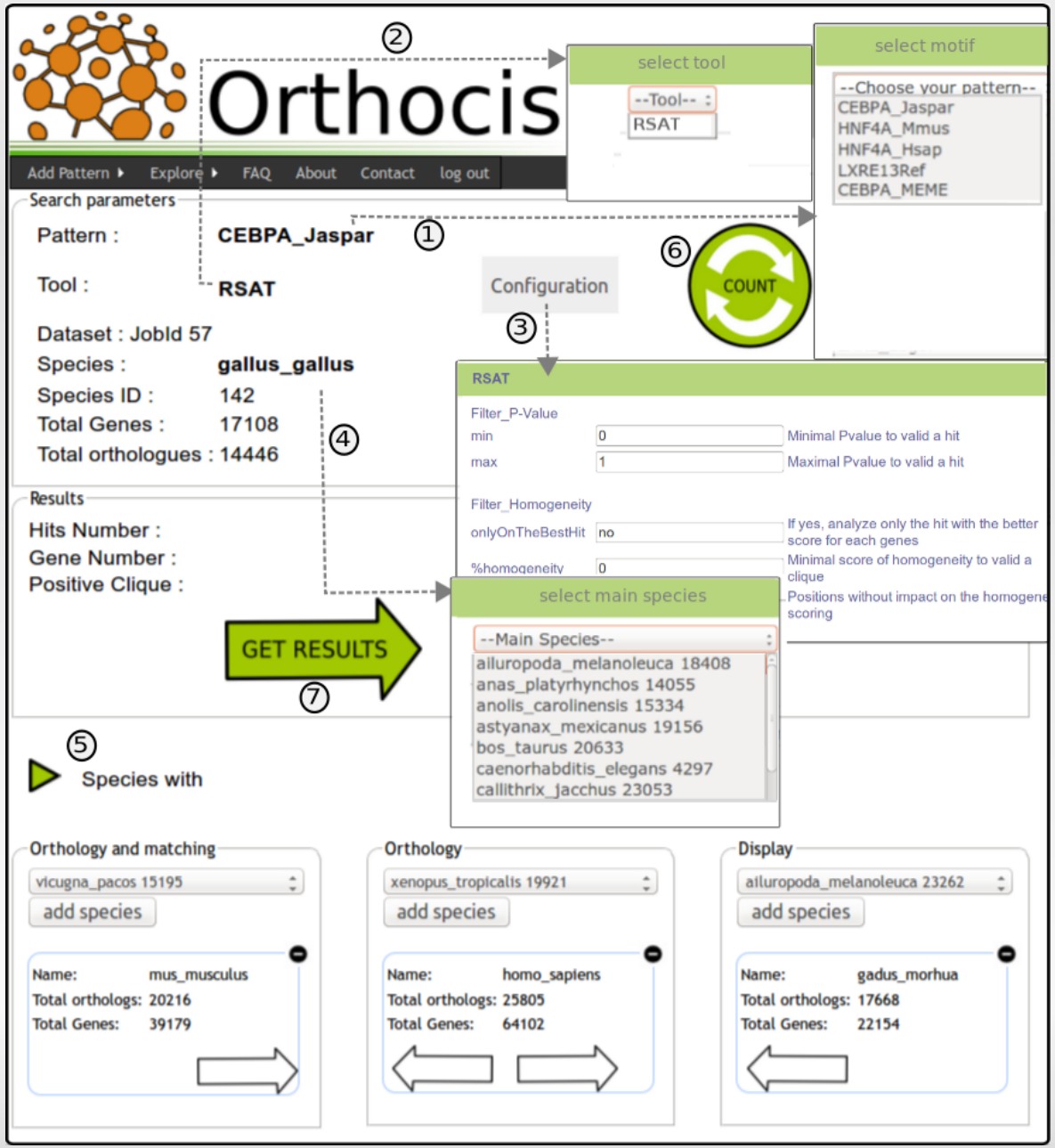
Orthocis, une base de données 2/2

- **Quels gènes d'Ensembl sont intégrés à la base ?**
  - Ceux possédant au moins 1 gène orthologue (one2one)
  - Mais qui ne possèdent pas de paralogues (one2many)





Orthocis, une interface web



Présentation des filtres 1/2

- Objectif : Limiter l'avalanche de faux positifs lors des *pattern matching* de TFBS plein génome

  FT LXR chez l'humain = 86 060 hits pour 25 806 gènes

- Orthocis permet de filtrer les résultats de *pattern matching* selon 3 filtres différents :

  - un filtre d'orthologie

  - un filtre de Pvaleur

  - un filtre d'homogénéité

Présentation des filtres 2/2

- Objectif : Limiter l'avalanche de faux positifs lors des *pattern matching* de TFBS plein génome

  FT LXR chez l'humain = 86 060 hits pour 25 806 gènes

- Orthocis permet de filtrer les résultats de *pattern matching* selon 3 filtres différents :

  - un filtre d'orthologie

  - un filtre de Pvaleur

  - un filtre d'homogénéité

- Exemple au travers du cas d'étude : recherche de nouveaux gènes candidats à une régulation par le FT LXR

Présentation de LXR 1/5

- **Données initiales du TFBS LXR**
  - 13 références validées biologiquement

1. Cyp7alpha1	Souris	TGAACTtgggTGACCA
2. Cyp7Alpha1	Rat	TGAACTtgagTGACCA
3. FASN	Souris	TGACCGgtagTAACCC
4. FASN	Rat	TGACCGgtagTAACCC
5. FASN	Poule	TGACCTgtggTAACCT
6. FASN	Humain	TGACCGgcagTAACCC
7. LPCAT3	Humain	CGACCGggagTAACCT
8. LPCAT3	Souris	CGACCGagagTAACCT
9. LPCAT3	Rat	CGACCGagagTAACCT
10. LPCAT3	Poule	TGCCCGcgagTAACCC
11. CETP	Humain	TGCCCGacaaTGACCC
12. CYP51a	Humain	TGACCTcaggTGATCC
13. SCD1	Souris	TGACCAcaggTAACCT

Présentation de LXR 2/5

- **Données initiales du TFBS LXR**
  - 13 références validées biologiquement

1. Cyp7alpha1	Souris	TGAACTtgggTGACCA
2. Cyp7Alpha1	Rat	TGAACTtgagTGACCA
3. FASN	Souris	TGACCGgtagTAACCC
4. FASN	Rat	TGACCGgtagTAACCC
5. FASN	Poule	TGACCTgtggTAACCT
6. FASN	Humain	TGACCGgcagTAACCC
7. LPCAT3	Humain	CGACCGggagTAACCT
8. LPCAT3	Souris	CGACCGagagTAACCT
9. LPCAT3	Rat	CGACCGagagTAACCT
10. LPCAT3	Poule	TGCCCGcgagTAACCC
11. CETP	Humain	TGCCCGacaaTGACCC
12. CYP51a	Humain	TGACCTcaggTGATCC
13. SCD1	Souris	TGACCAcaggTAACCT

Motif DR4

Présentation de LXR 3/5

- **Données initiales du TFBS LXR**
  - 13 références validées biologiquement
  - Matrice poids-position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0	11	2	0	1	3,25	3,25	3,25	3,25	0	9	13	0	0	2
C	3	0	2	11	13	0	3,25	3,25	3,25	3,25	0	0	0	12	13	6
G	0	13	0	0	0	8	3,25	3,25	3,25	3,25	0	4	0	0	0	0
T	10	0	0	0	0	4	3,25	3,25	3,25	3,25	13	0	0	1	0	5

Présentation de LXR 4/5

- **Données initiales du TFBS LXR**
  - 13 références validées biologiquement
  - Matrice poids-position
  - Liste de 840 gènes fortement différentiellement exprimés (DE) entre des souris sauvages et des souris LXR KO

    - Gènes DE : cibles discutables pour une régulation directe par LXR
    - Mais plus grande probabilité d'être ciblés par LXR malgré tout

Les gènes DE ont été utilisés comme « cibles positives » pour les tests suivants

Cas d'étude : LXR

- **Données initiales du TFBS LXR**
  - 13 références validées biologiquement
  - Matrice poids-position
  - Liste de 840 gènes fortement différentiellement exprimés (DE) entre des souris sauvages et des souris LXR KO

- *Recherche du motif LXR chez l'humain (25 806 gènes)*
  23 459 gènes avec au moins un hit (i.e. ~91 %)
  - dont 840 de liste DE (i.e. ~3,5 % d'enrichissement)

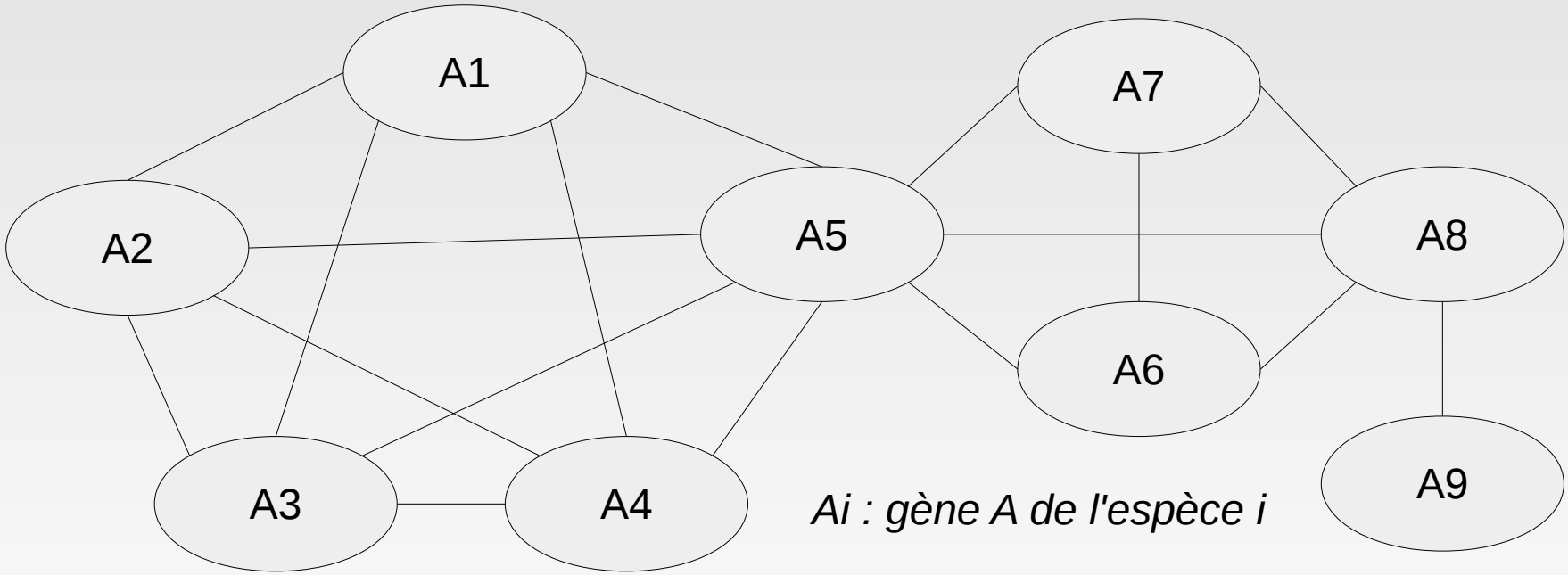
Filtre d'orthologie 1/5

- ***Filtrer selon l'orthologie***

  « Vérifier la conservation du gène au travers des espèces »

Filtre d'orthologie 2/5

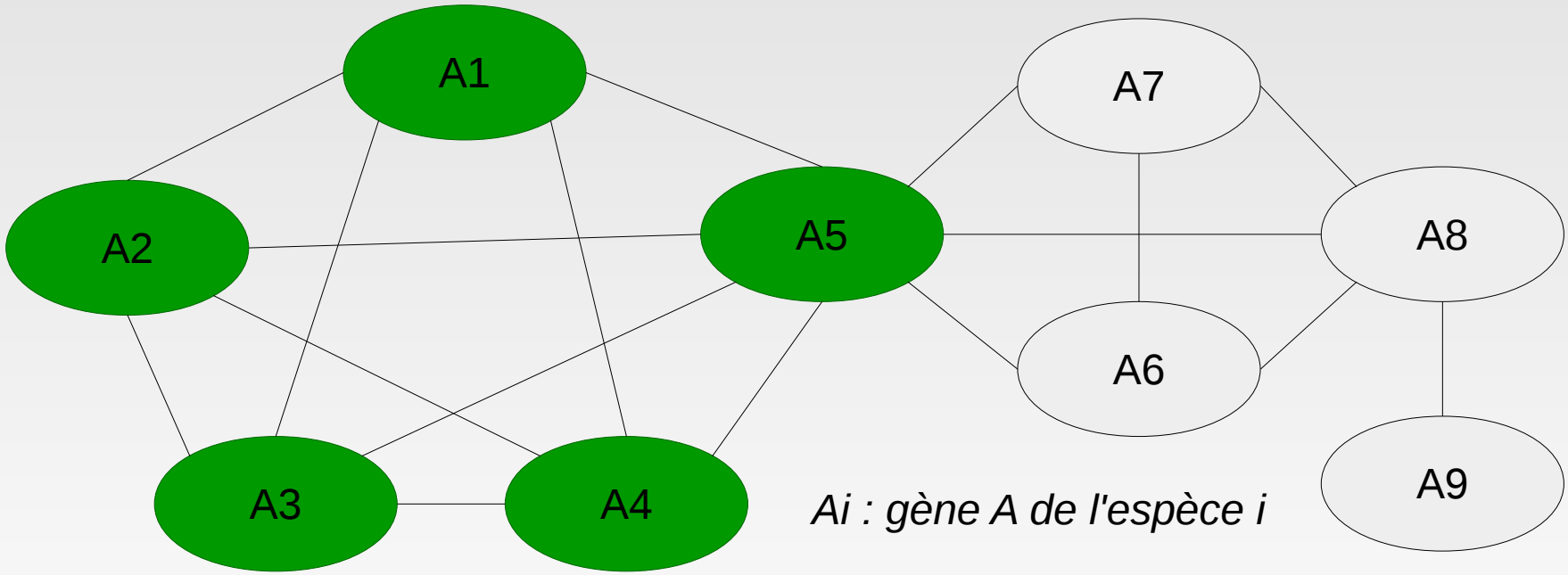
- ***Filtrer selon l'orthologie***
  - Utilisateur : sélection d'un set d'espèces
  - Obtention des gènes formant une clique d'orthologie



Composante connexe d'orthologie des gènes A

Filtre d'orthologie 3/5

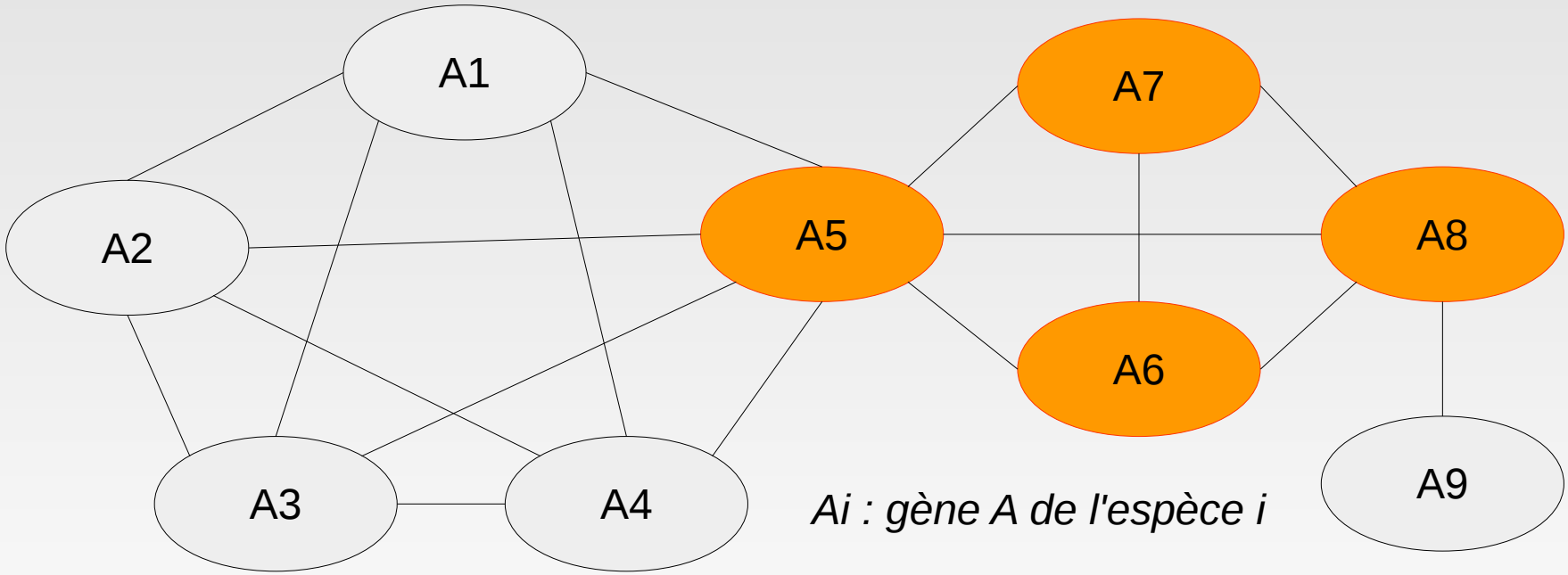
- ***Filtrer selon l'orthologie***
  - Utilisateur : sélection d'un set d'espèces
  - Obtention des gènes formant une clique d'orthologie



Les espèces 1-2-3-4-5 forment une clique pour le gène A
Ils sont orthologues 2 à 2 avec tous les autres

Filtre d'orthologie 4/5

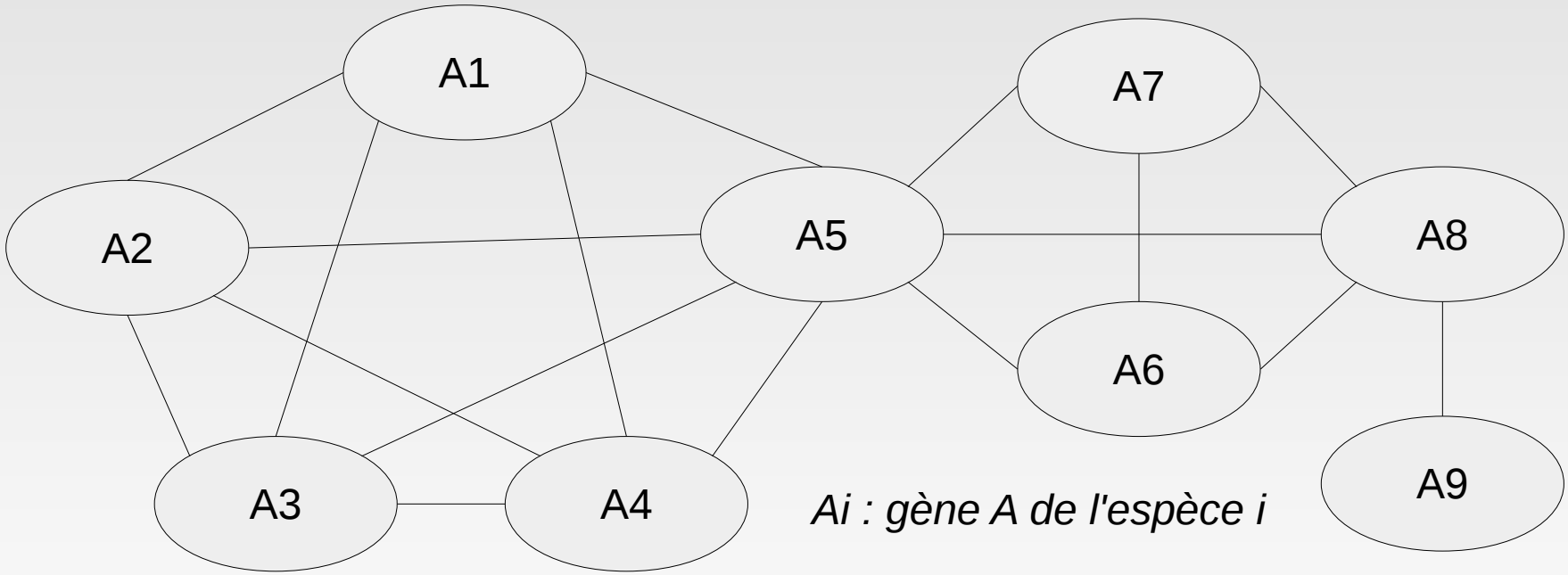
- ***Filtrer selon l'orthologie***
  - Utilisateur : sélection d'un set d'espèces
  - Obtention des gènes formant une clique d'orthologie



Les espèces 5-6-7-8 forment une clique pour le gène A
Ils sont orthologues 2 à 2 avec tous les autres

Filtre d'orthologie 5/5

- ***Filtrer selon l'orthologie***
  - Utilisateur : sélection d'un set d'espèces
  - Obtention des gènes formant une clique d'orthologie



Filtre d'orthologie : ne sélectionner un gène que s'il est conservé dans les espèces sélectionnées, c'est-à-dire s'il forme une clique avec les gènes orthologues du set d'espèce (principe de conservation au cours de l'évolution)

Cas d'étude : LXR

- ***Filtrer selon l'orthologie***

- *Recherche du motif LXR chez l'humain (25 806 gènes)*
  23 459 gènes avec au moins un hit (i.e. ~91 %)
  - dont 840 de liste DE (i.e. ~3,5 % d'enrichissement)

- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*
  3 902 gènes avec au moins un hit (i.e. ~44,8 %)
  - dont 213 de liste DE (i.e. ~5,46 % d'enrichissement)

Filtre de Pvaleur 1/2

- ***Filtrer selon la Pvaleur***

  « Vérifier que les hits ressemblent au consensus du TFBS »

Filtre de Pvaleur 2/2

- ***Filtrer selon la Pvaleur***
  - La Pvaleur d'un ensemble de hits d'un groupe de gènes orthologues est égale à la plus mauvaise Pvaleur présente dans l'ensemble

    Gène A, espèce 1 : 5,5e-07
    Gene A, espèce 2 : 2,7e-05
    Gene A, espèce 3 : 7,8e-05
    } Clique Gène A : 7,8e-05

  - Utilisateur : sélection d'une valeur seuil (ex : plus mauvaise valeur des hits de référence)
  - Obtention des cliques passant ce seuil

Filtre Pvaleur : ressembler suffisamment à la cible

Cas d'étude : LXR

- ***Filtrer selon la Pvaleur***
  - *Exemple de l'étude de LXR*

1. Cyp7alpha1	Souris	TGAACTtgggTGACCA	2e-05
2. Cyp7Alpha1	Rat	TGAACTtgagTGACCA	**2,1e-05**
3. FASN	Souris	TGACCGgtagTAACCC	8,4e-09
4. FASN	Rat	TGACCGgtagTAACCC	1,1e-08
5. FASN	Poule	TGACCTgtggTAACCT	4,8e-07
6. FASN	Humain	TGACCGgcagTAACCC	2,4e-08
7. LPCAT3	Humain	CGACCGggagTAACCT	2,4e-08
8. LPCAT3	Souris	CGACCGagagTAACCT	1,8e-08
9. LPCAT3	Rat	CGACCGagagTAACCT	2,2e-08
10. LPCAT3	Poule	TGCCCGcgagTAACCC	7,3e-08
11. CETP	Humain	TGCCCGacaaTGACCC	7,8e-07
12. CYP51a	Humain	TGACCTcaggTGATCC	8,3e-06
13. SCD1	Souris	TGACCAcaggTAACCT	2,4e-06

Cas d'étude : LXR

- ***Filtrer selon la Pvaleur***
- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*

  3 902 gènes avec au moins un hit (i.e. ~44,8 %)

  - dont 213 de liste DE (i.e. ~5,46 % d'enrichissement)

- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*
- *Filtre de Pvaleur : 2,1e-05*

  288 gènes avec au moins un hit (i.e. ~3,31 %)

  - dont 16 de liste DE (i.e. ~5,56 % d'enrichissement)

Filtre d'homogénéité 1/4

- ***Filtrer selon l'homogénéité***

« Vérifier que les hits de gènes orthologues se ressemblent entre eux »

Filtre d'homogénéité 2/4

- ***Filtrer selon l'homogénéité***
  - Utilisateur : sélection d'un score minimum d'homogénéité
  - Obtention des cliques dont les hits intra-cliques se ressemblent entre eux

Filtre d'homogénéité 3/4

- ***Filtrer selon l'homogénéité***

- *Calcul d'homogénéité*

  Espèce 1 : `TGACC`**`T`**`GGAGT`**`G`**`ACCC`
  Espèce 2 : `TGACCGGTAGT`**`G`**`ACC`**`T`**
  Espèce 3 : `TGACCG`**`AG`**`A`**`T`**`TCACCC`
  Espèce 4 : `TGACC`**`T`**`GTAGTCACC`**`T`**
  Ensemble `44444---4-4-444-` : Score 10/16

  - Score = nombre de colonnes ayant la même valeur

Filtre d'homogénéité : renforcement de l'hypothèse de conservation des TFBS au cours de l'évolution

Filtre d'homogénéité 4/4

- ***Filtrer selon l'homogénéité***

  - Si plusieurs hits sont possibles pour un gène, toute la combinatoire est testée afin d'obtenir l'ensemble le plus homogène
  - La clique des hits les plus homogènes peut être différente de la clique des hits avec la meilleure Pvaleur

Filtre d'homogénéité : renforcement de l'hypothèse de conservation des TFBS au cours de l'évolution

Cas d'étude : LXR

- ***Filtrer selon l'homogénéité***
- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*
- *Filtre de Pvaleur : 2,1e-05*

  288 gènes avec au moins un hit (i.e. ~3,31 %)

  - dont 16 de liste DE (i.e. ~5,56 % d'enrichissement)

- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*
- *Filtre de Pvaleur : 2,1e-05*
- *Filtre sur l'homogénéité : 75 % (12/16)*

  2 gènes avec au moins un hit (i.e. ~0,02 %)

  - dont 2 de liste DE (i.e. 100 % d'enrichissement)

Cas d'étude : LXR

- ***Filtrer selon l'homogénéité***

- *Recherche du motif LXR*
- *Chez 6 espèces (Humain, chimpanzé, macaque, souris, rat, poule) (8 698 gènes)*
- *Filtre de Pvaleur : 2,1e-05*
- *Filtre sur l'homogénéité : 75 % (12/16)*

  2 gènes avec au moins un hit (i.e. ~0,02 %)

  - dont 2 de liste DE (i.e. 100 % d'enrichissement)

1 référence retrouvée (LPCAT3)
+ 1 nouveau candidat identifié
→ Présomption très forte de sa validité

Cas d'étude LXR : conclusions

La multiplication des filtres a permis d'identifier des candidats probables à une régulation par LXR

Le rappel est faible (2 gènes) mais les candidats sont « solides » et à quantité « humaine », permettant leur étude : résultat très intéressant pour les biologistes

Il doit être possible d'améliorer le rappel tout en restant dans des quantités raisonnables (curseur des filtres, modification des contraintes)

Orthocis : conclusions

- **De façon plus générale** :
  - Orthocis se base sur le principe de conservation des TFBS au sein de l'évolution pour aider à leur détection
  - Orthocis permet à l'utilisateur :
    - De soumettre un FT et de stocker les résultats d'un pattern matching permissif
    - D'utiliser plusieurs leviers pour filtrer les résultats
      - Le score des hits
      - La conservation inter-espèce
      - L'homogénéité des TFBS obtenus
  - Les perspectives d'améliorations sont :
    - La prise en compte de l'orthologie one2many
    - Le couplage à l'exploitation de données ChIP-Seq

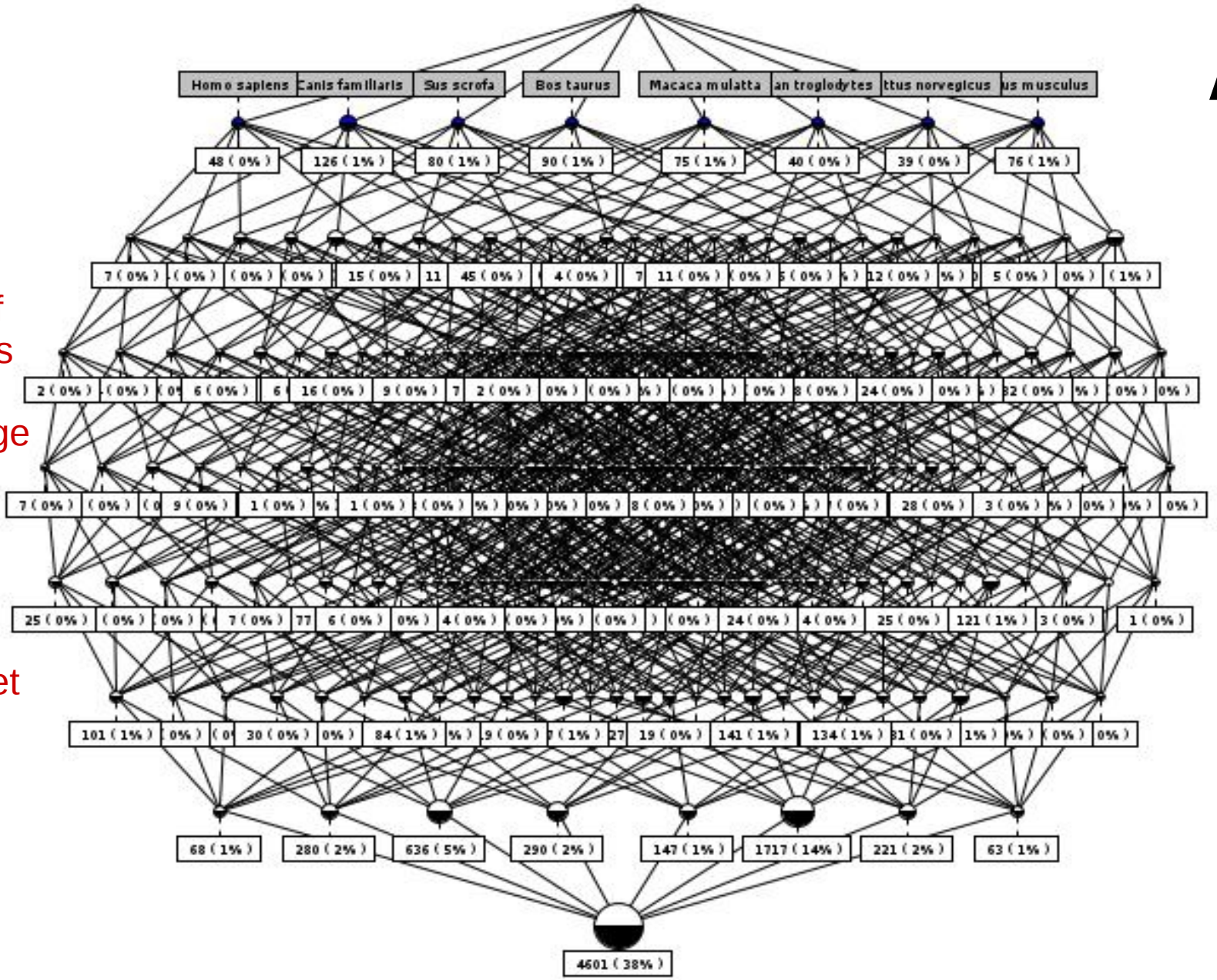
Merci de votre attention

Lattice structure

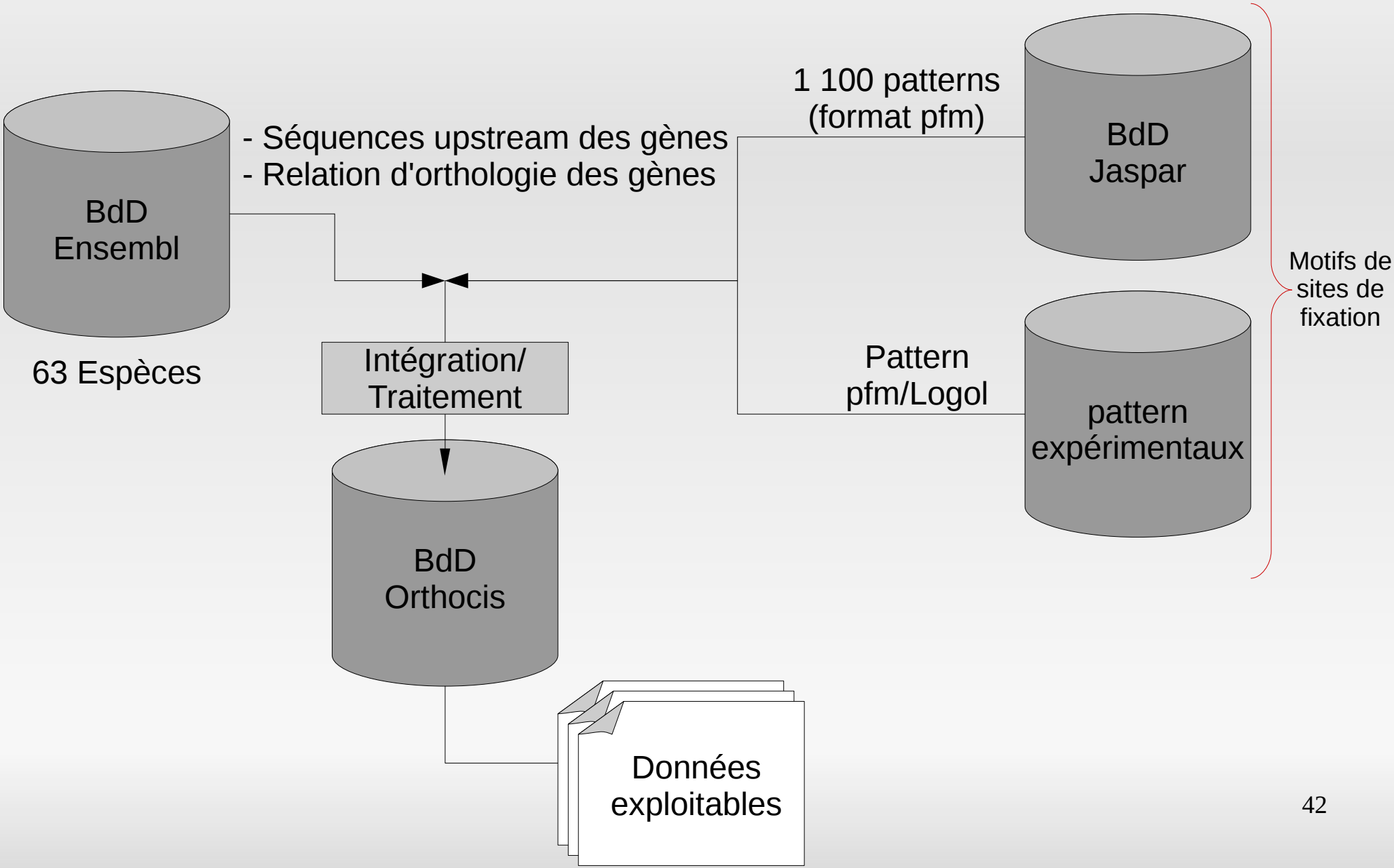
- For a references species

- For a set of other species

=> Knowledge of specific orthologous genes for each combination of species set



Application 2 : recherche de FT
La base de données Orthocis



Application 2 : recherche de FT
Structure mise en place