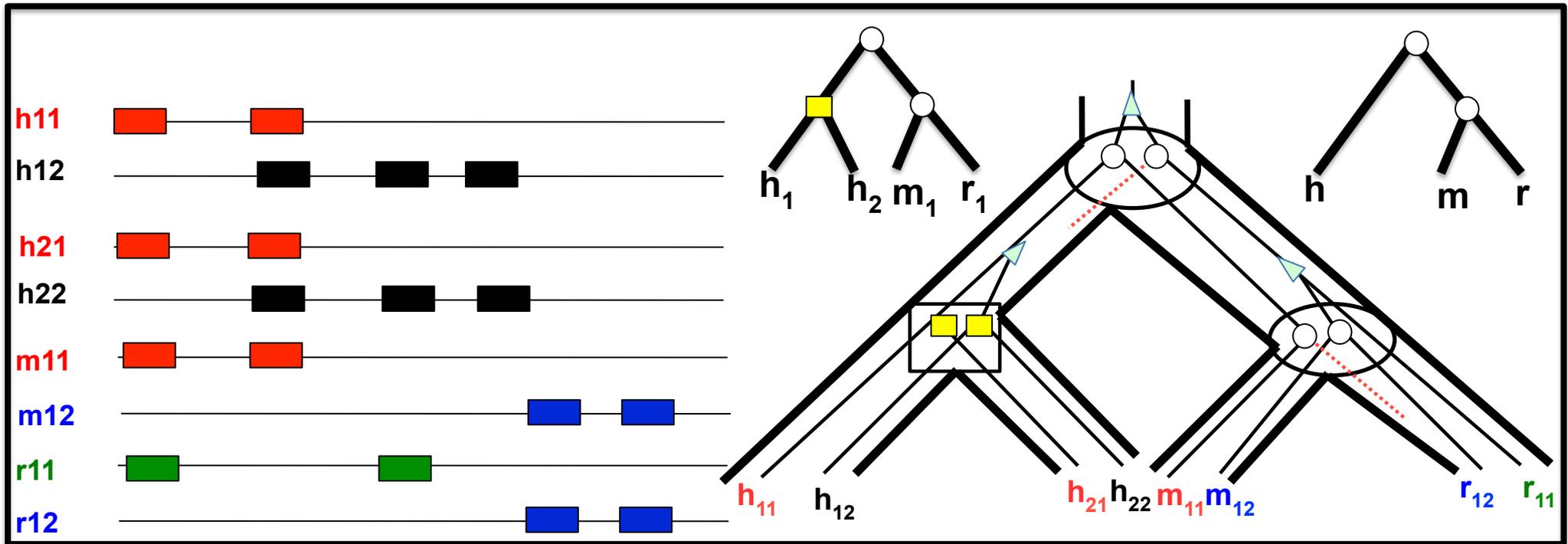


Les secrets complexes de notre ADN



Aïda Ouangraoua
Laboratoire CoBIUS

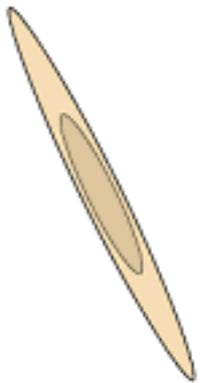
Département d'informatique, Faculté des sciences

Complexité définie par les niveaux d'organisation

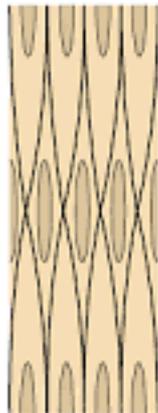
❑ Organisme

- ❑ Des appareils et systèmes (fonctions complexes communes)
 - ❑ Des organes (fonctions physiologiques)
 - ❑ Différents types de tissus (ensembles fonctionnels)
 - ❑ Ensemble de cellules semblables

cellule



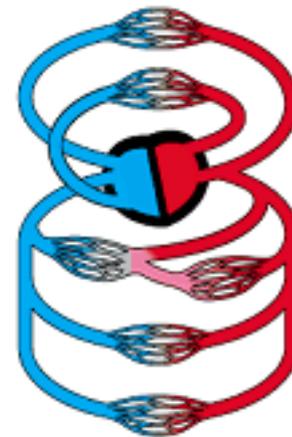
tissu



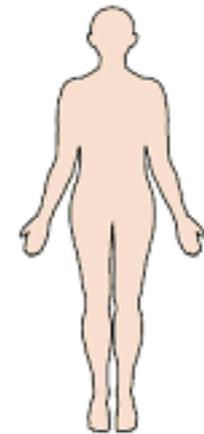
organe



système



organisme



<http://www.colscol.com/body-system/>

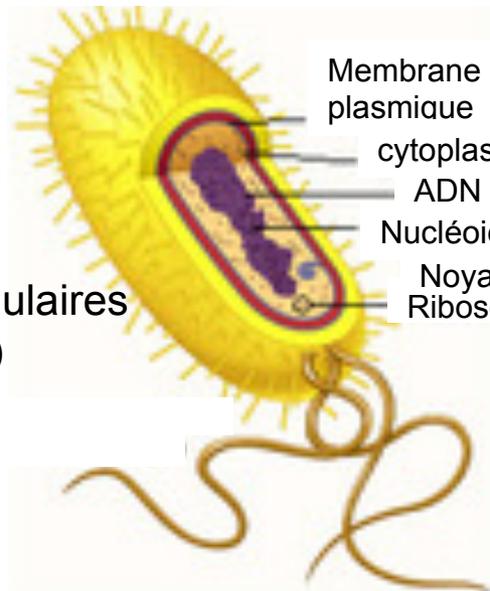
Complexité définie par les niveaux d'organisation

❑ Cellule

- ❑ Composante de base de tous les organismes vivants
- ❑ Plus petite unité capable de fonctionner de façon autonome
- ❑ Stocke et transmet l'information génétique (ADN)
- ❑ Réalise les actions chimiques nécessaires à la vie

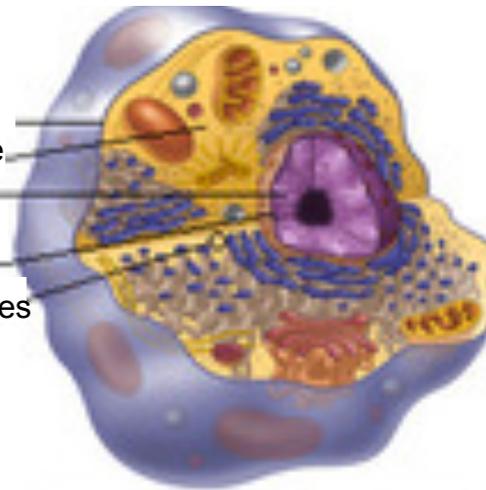
Procaryote (moins complexe)

- ◆ Pas de noyau
- ◆ Organismes unicellulaires (bactéries, archées)



Eucaryote (plus complexe)

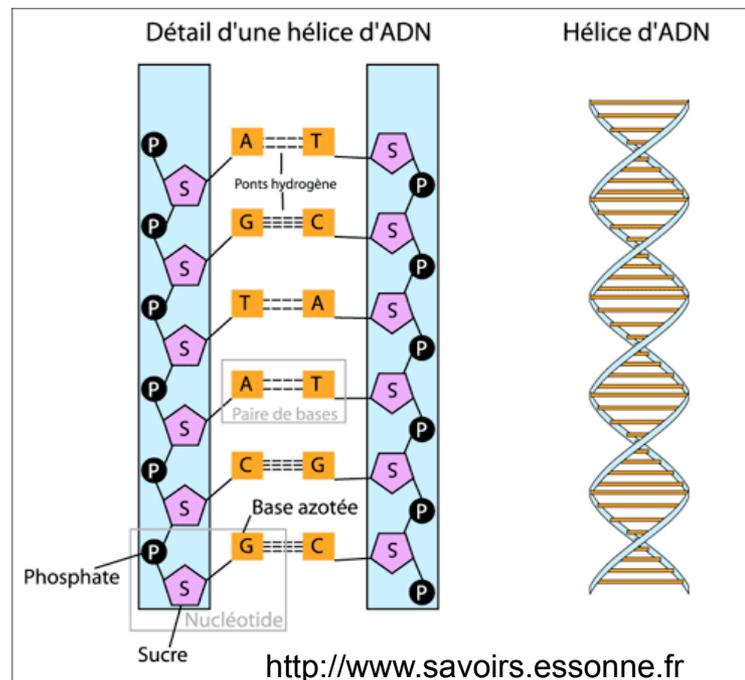
- ◆ Noyau délimité par une membrane
- ◆ Organismes uni et pluri-cellulaires



<http://cells-breannaolivia.weebly.com>

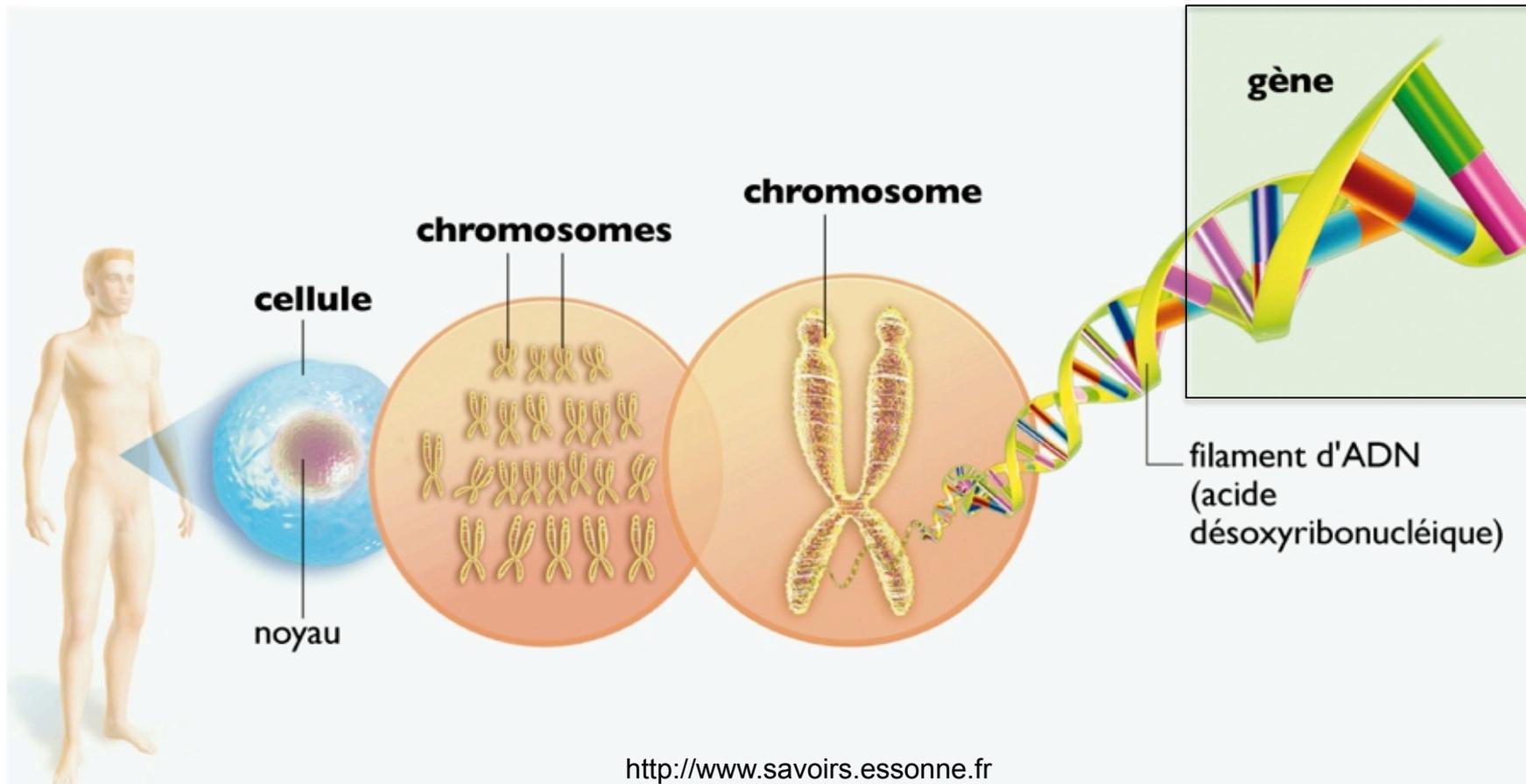
Génome

- ❑ Information génétique stockée dans la cellule sous forme d'ADN
- ❑ ADN (Acide DésoxyriboNucléique)
 - ❑ Deux chaînes de nucléotides formant une double hélice
 - ❑ Un nucléotide : un sucre, un groupe phosphate et une base (A, C, G ou T) caractérisant le nucléotide
 - ❑ Complémentarité entre les bases des deux chaînes (A-T, G-C)



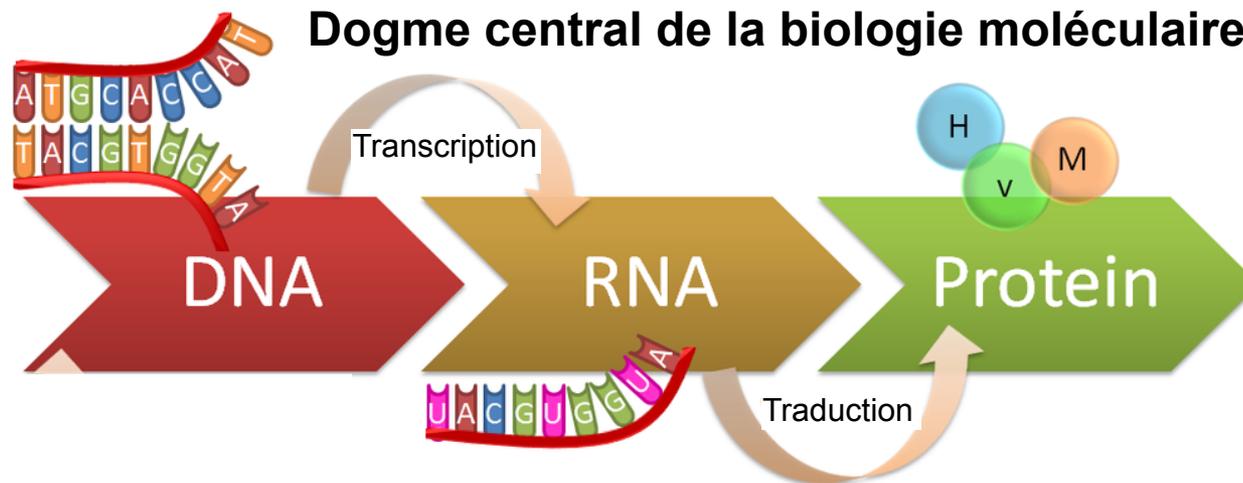
Génome

- ❑ Gènes : portions fonctionnelles de l'ADN



Génome

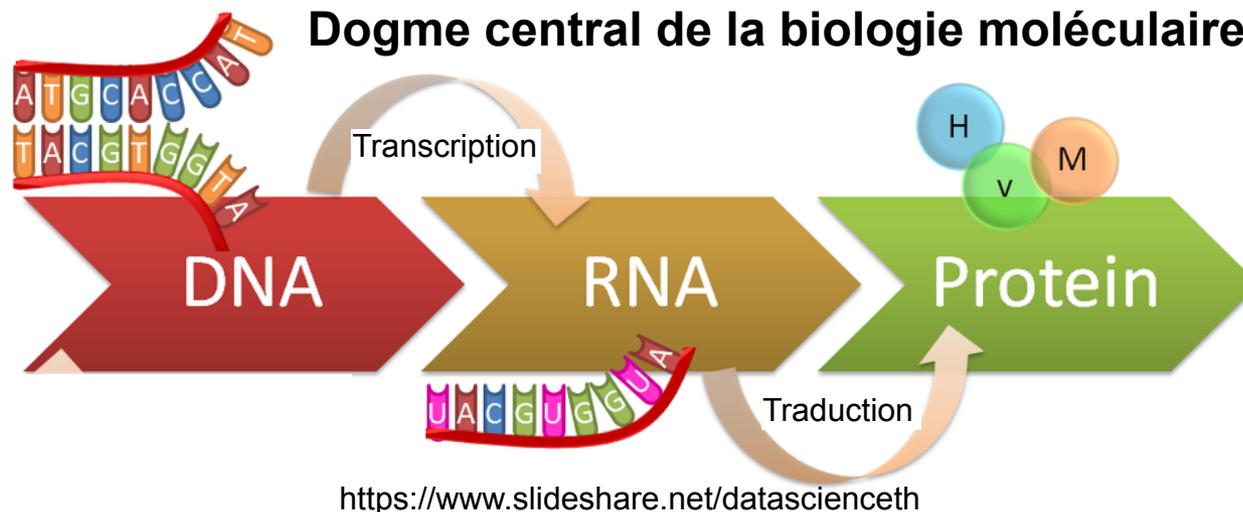
- ❑ Gènes : portions à partir desquelles sont générés des produits fonctionnels
 - ❑ ARN : Acide RiboNucléique (similaire à l'ADN mais en simple brin et T → U)
 - ❑ Protéines : chaînes d'acides aminés



<https://www.slideshare.net/datascienceth>

Génome

- ❑ Gènes : portions à partir desquelles sont générés des produits fonctionnels
 - ❑ ARN : Acide RiboNucléique (similaire à l'ADN mais en simple brin et T → U)
 - ❑ Protéines : chaînes d'acides aminés



- ❑ Gènes codants : servent à produire des ARN dit messagers à partir desquels sont produites des protéines
- ❑ Gènes non-codants : ne servent à produire que des ARN dit non-codants

Complexité des organismes

- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?
 - ❑ la taille de l'organisme (nombre et volume des cellules) ?
 - ◆ Non, il existe de grands organismes simples et des petits organismes complexes

Grandes éponges marines



Microscopiques rotifères marins



Complexité des organismes

- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?
 - ❑ la taille de l'organisme (nombre et volume des cellules) ?
 - ◆ Non, il existe de grands organismes simples et des petits organismes complexes

Grandes éponges marines



Microscopiques rotifères marins



- ❑ la taille du génome (longueur de l'ADN) ?

◆ Non : **Amibe unicellulaire (*P. Dubium*) (775 Gb)**

Génome humain (3Gb)

Complexité des organismes

- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?

- ❑ la taille de l'organisme (nombre et volume des cellules) ?

- ◆ Non, il existe de grands organismes simples et des petits organismes complexes

Grandes éponges marines



Microscopiques rotifères marins



- ❑ la taille du génome (longueur de l'ADN) ?

- ◆ Non : **Amibe unicellulaire (*P. Dubium*) (775 Gb)**

Génome humain (3Gb)

- ❑ la proportion en gènes ?

- ◆ Non : **Bactérie unicellulaire (*E. Coli*) (90% codant)**

Génome humain (3%)

Complexité des organismes

- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?

- ❑ la taille de l'organisme (nombre et volume des cellules) ?

◆ Non, il existe de grands organismes simples et des petits organismes complexes

Grandes éponges marines



Microscopiques rotifères marins



- ❑ la taille du génome (longueur de l'ADN) ?

◆ Non : **Amibe unicellulaire (*P. Dubium*) (775 Gb)**

Génome humain (3Gb)

- ❑ la proportion en gènes ?

◆ Non : **Bactérie unicellulaire (*E. Coli*) (90% codant)**

Génome humain (3%)

- ❑ le nombre de gènes ?

◆ Non : **Petit nématode (*C. Elegans*)**

20,222 codants

24,765 non-codants

Génome humain

20,338

22,521

Complexité des organismes

- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?

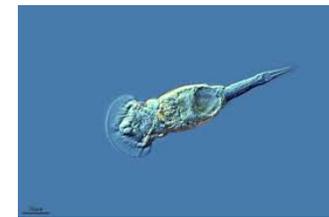
- ❑ la taille de l'organisme (nombre et volume des cellules) ?

- ◆ Non, il existe de grands organismes simples et des petits organismes complexes

Grandes éponges marines



Microscopiques rotifères marins



- ❑ la taille du génome (longueur de l'ADN) ?

- ◆ Non : **Amibe unicellulaire (*P. Dubium*) (775 Gb)**

Génome humain (3Gb)

- ❑ la proportion en gènes ?

- ◆ Non : **Bactérie unicellulaire (*E. Coli*) (90% codant)**

Génome humain (3%)

- ❑ le nombre de gènes ?

- ◆ Non : **Petit nématode (*C. Elegans*)**

20,222 codants

24,765 non-codants

61,109 transcrits ARN

Génome humain

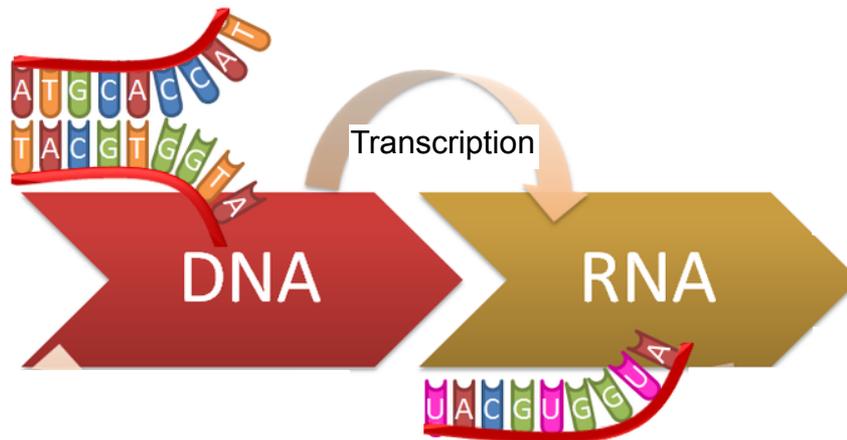
20,338

22,521

200,310

Complexité des organismes

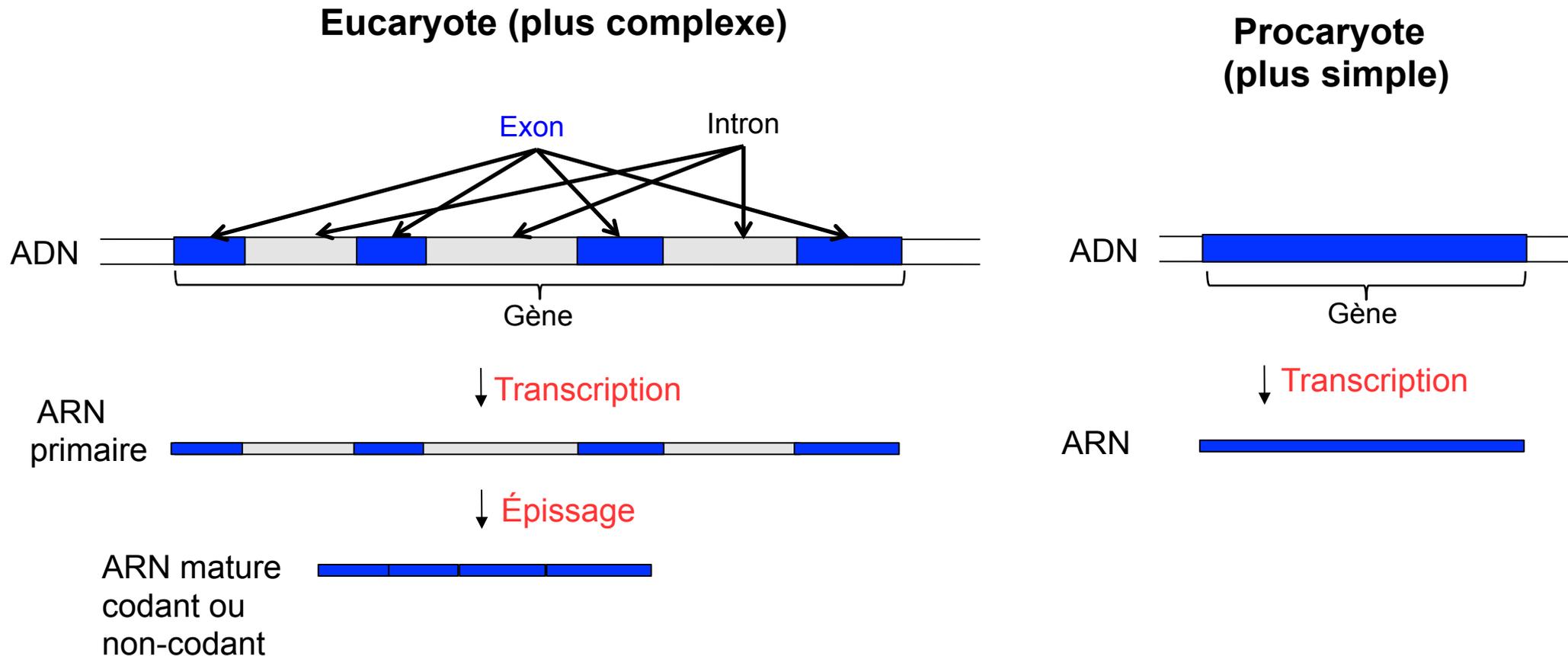
- ❑ Degré de spécialisation des cellules dans différents ensembles, tissus, organes, et systèmes
- ❑ Quels sont les facteurs qui expliquent la complexité ?
 - ❑ Le nombre de transcrits ARN différents produits à partir des gènes



- ❑ Processus de transcription différents chez les eucaryotes et les procaryotes

Complexité des organismes

- Processus de transcription différents chez les eucaryotes et les procaryotes

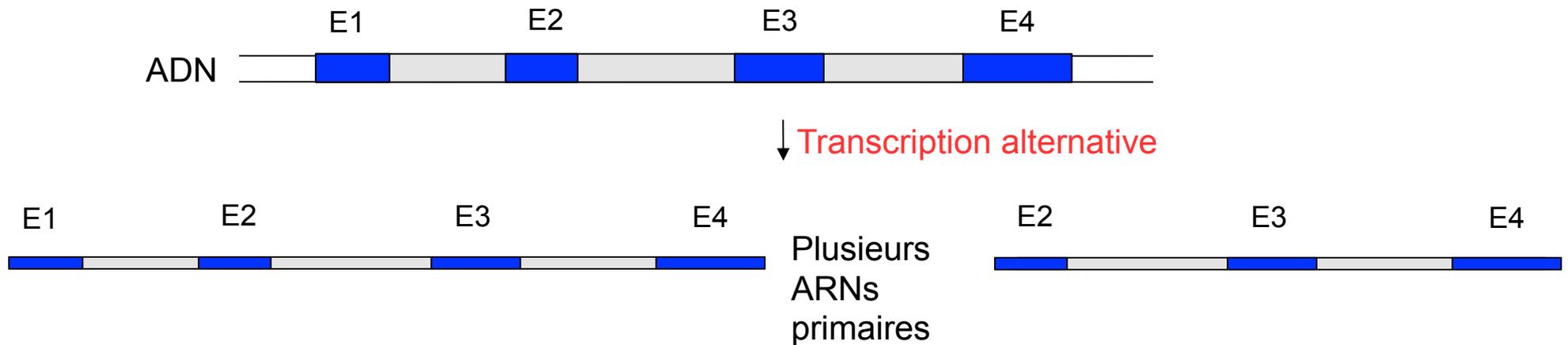


Traitement des gènes codants

Complexité des organismes

- La transcription alternative et l'épissage alternatif permettent de générer plusieurs ARNs différents chez les eucaryotes

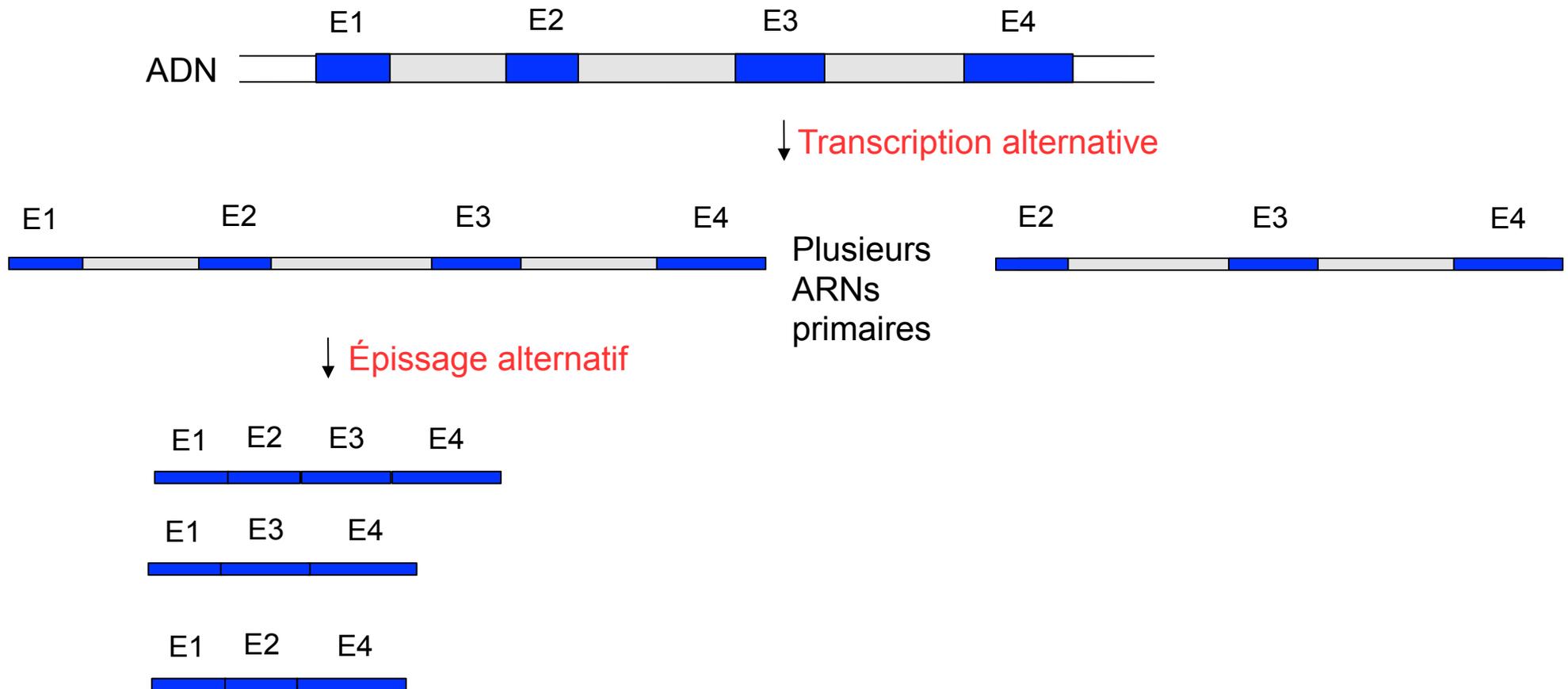
Transcription et épissage alternatif



Complexité des organismes

- La transcription alternative et l'épissage alternatif permettent de générer plusieurs ARNs différents chez les eucaryotes

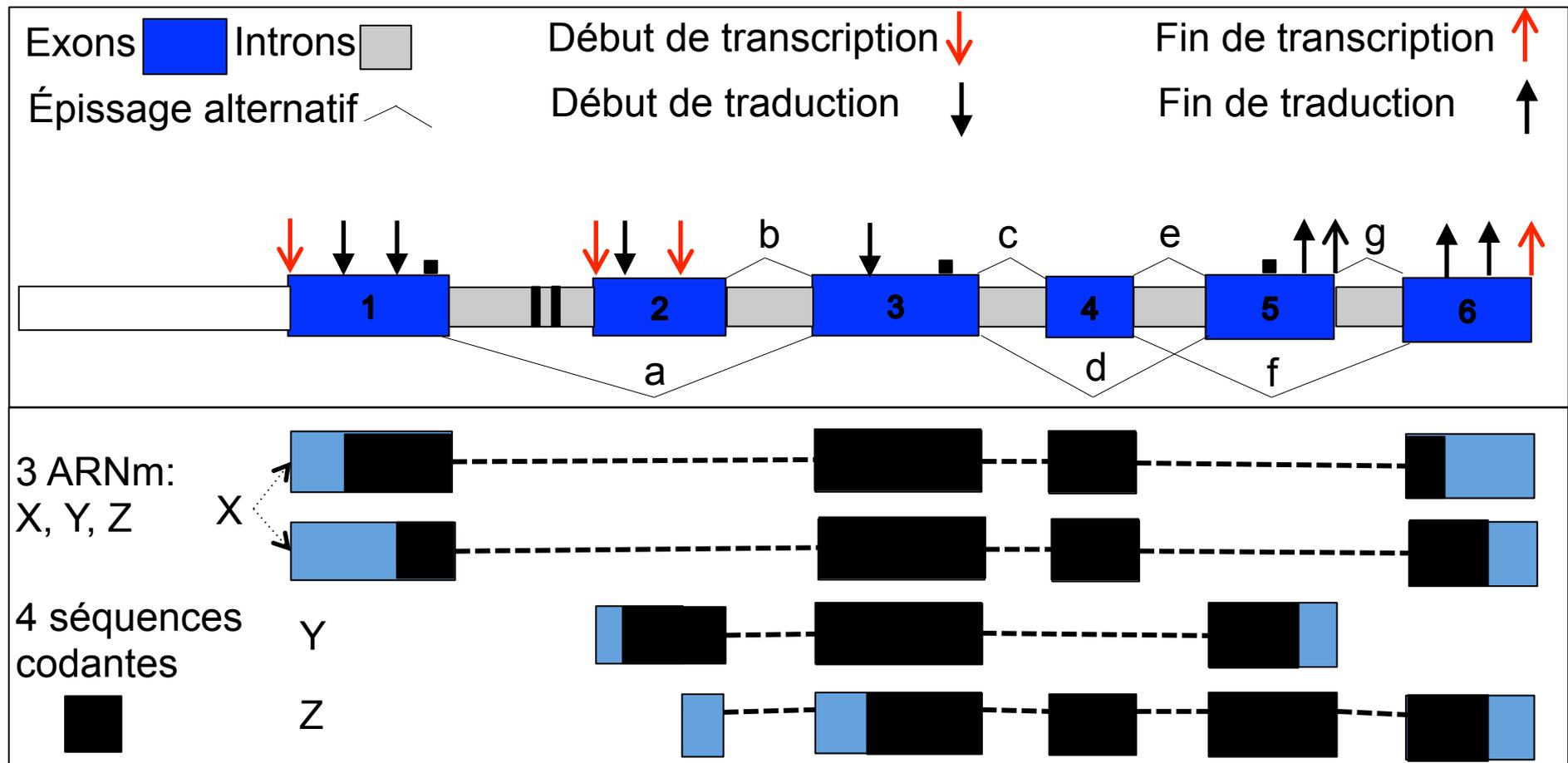
Transcription et épissage alternatif



- ◆ Plusieurs ARNs impliqués dans différentes fonctions

Complexité des organismes

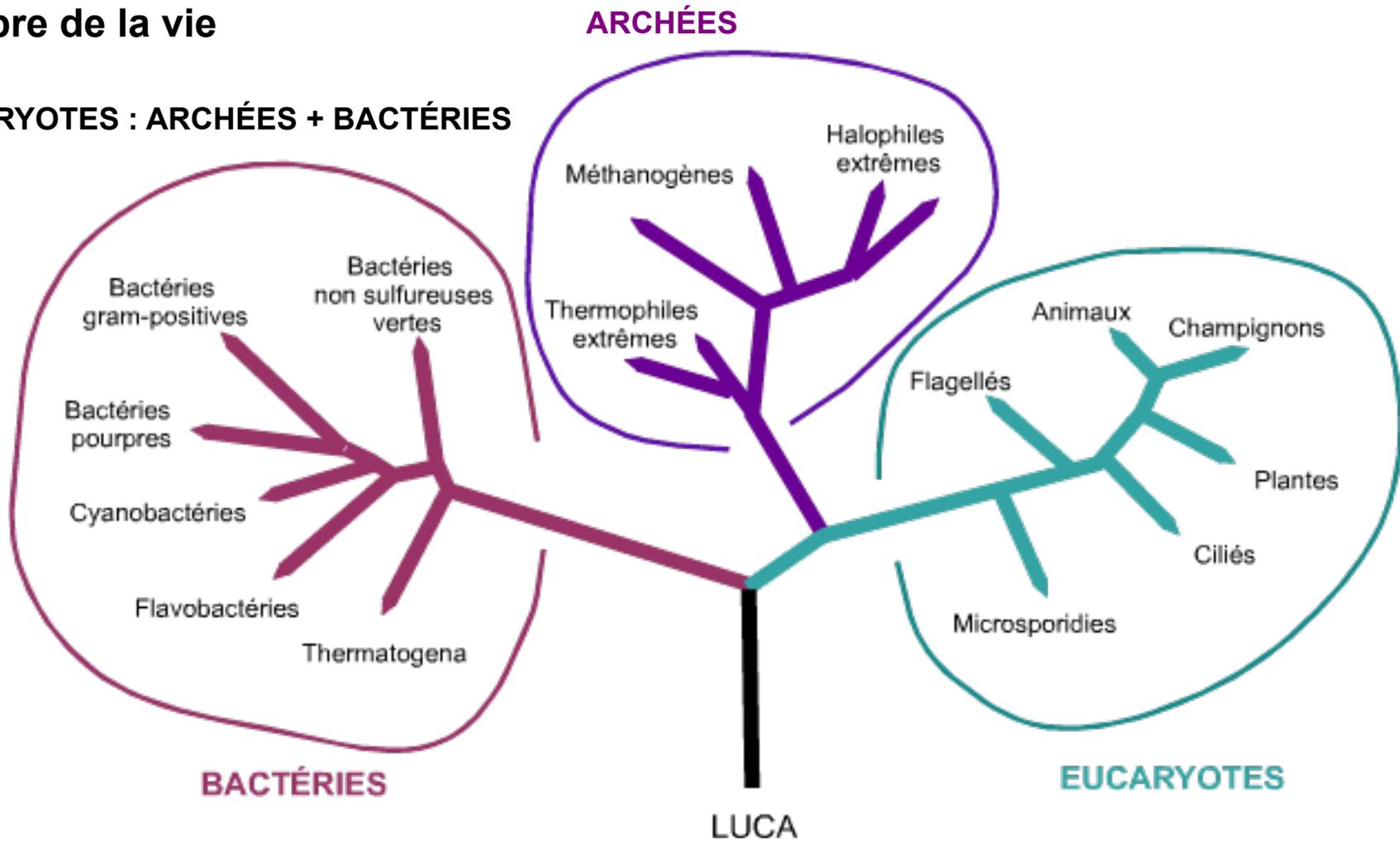
□ Structure d'un gène eucaryote



Évolution de la complexité

Arbre de la vie

PROCARYOTES : ARCHÉES + BACTÉRIES

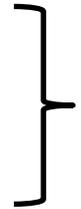


<http://www.evolution-biologique.org/>

Évolution des gènes

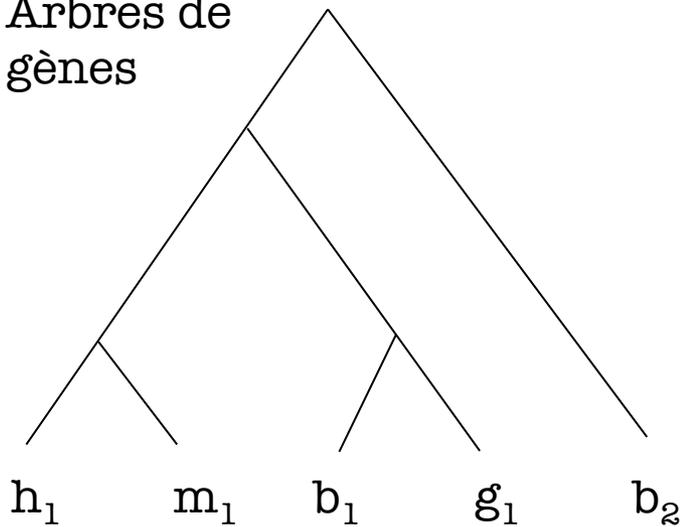
□ Considère les changements au niveau des gènes

- Gain de gènes
- Perte de gènes

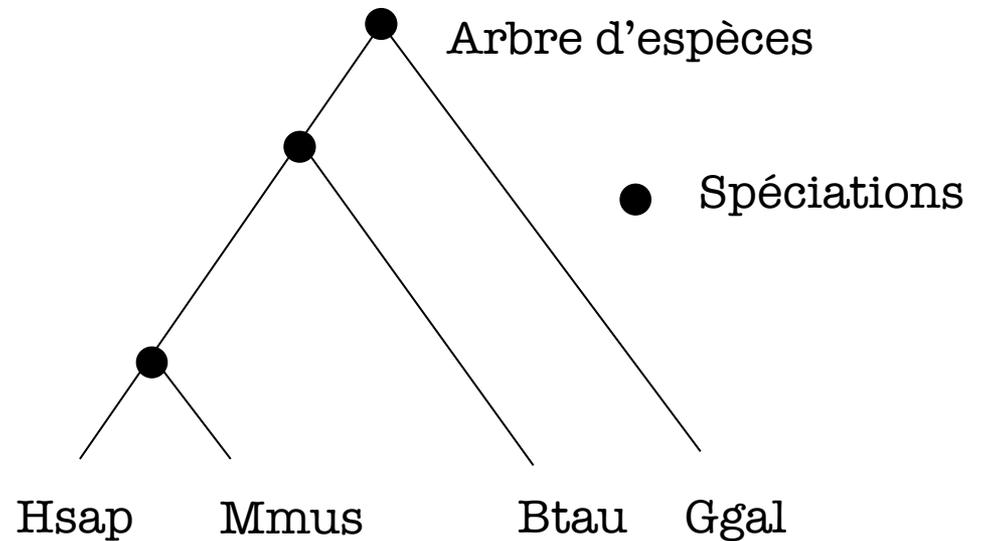


Expliquent la différence entre les arbres de gènes et les arbres d'espèces

Arbres de gènes



Arbre d'espèces



Ensemble de gènes homologues
Issus d'un même gène ancestral

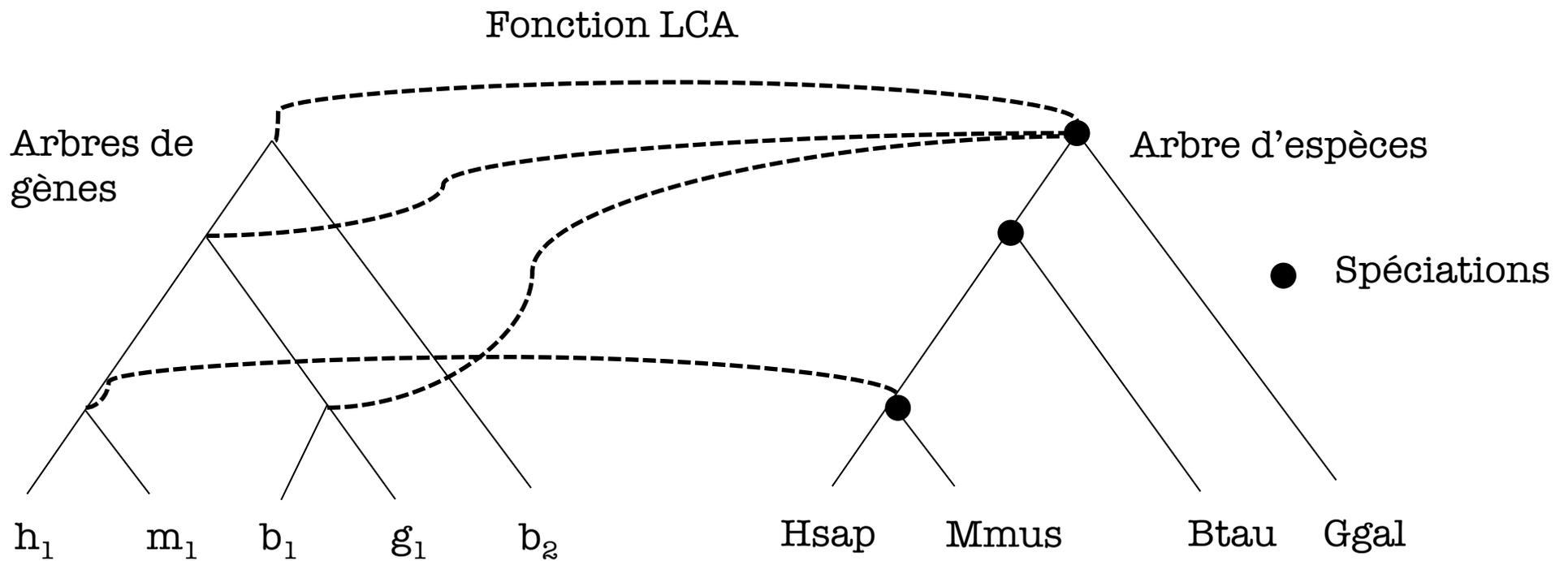
Évolution des gènes

□ Considère les changements au niveau des gènes

➤ Gain de gènes

➤ Perte de gènes

Expliquent la différence entre les arbres de gènes et les arbres d'espèces



Ensemble de gènes homologues
Issus d'un même gène ancestral

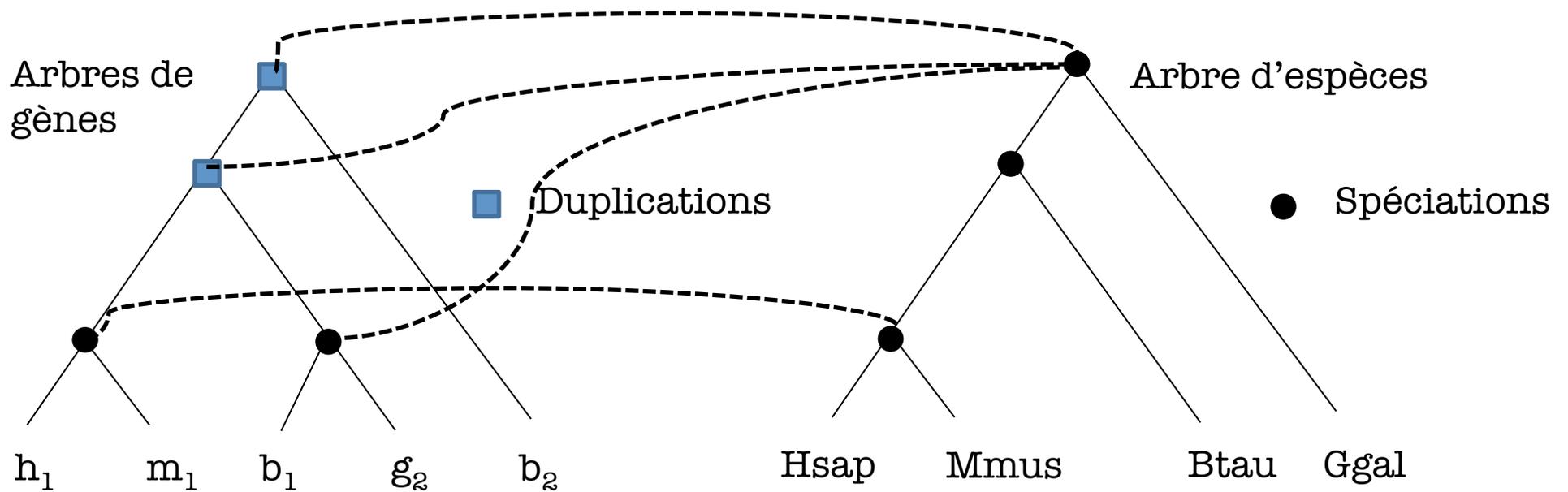
Évolution des gènes

□ Considère les changements au niveau des gènes

➤ Gain de gènes

➤ Perte de gènes

} Expliquent la différence entre les arbres de gènes et les arbres d'espèces

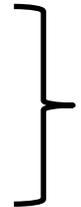


Ensemble de gènes homologues
Issus d'un même gène ancestral

Évolution des gènes

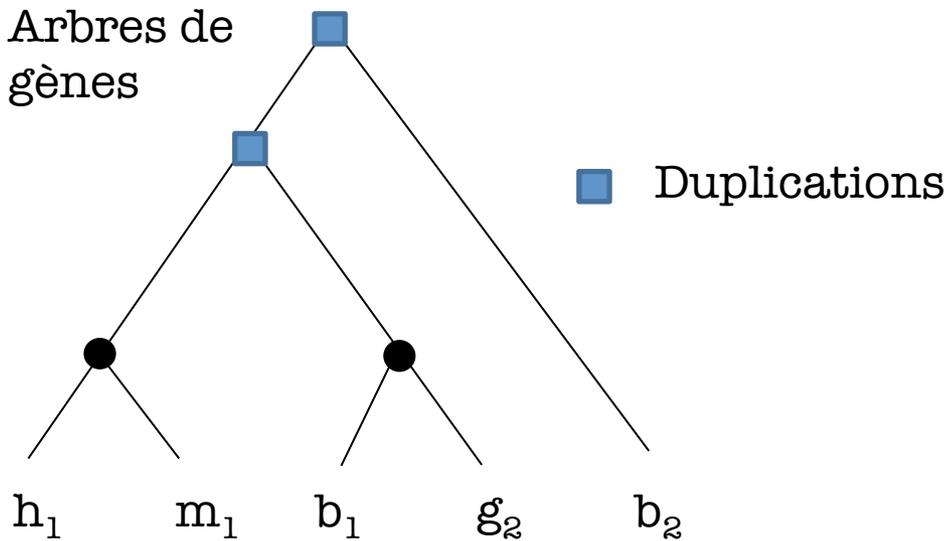
□ Considère les changements au niveau des gènes

- Gain de gènes
- Perte de gènes

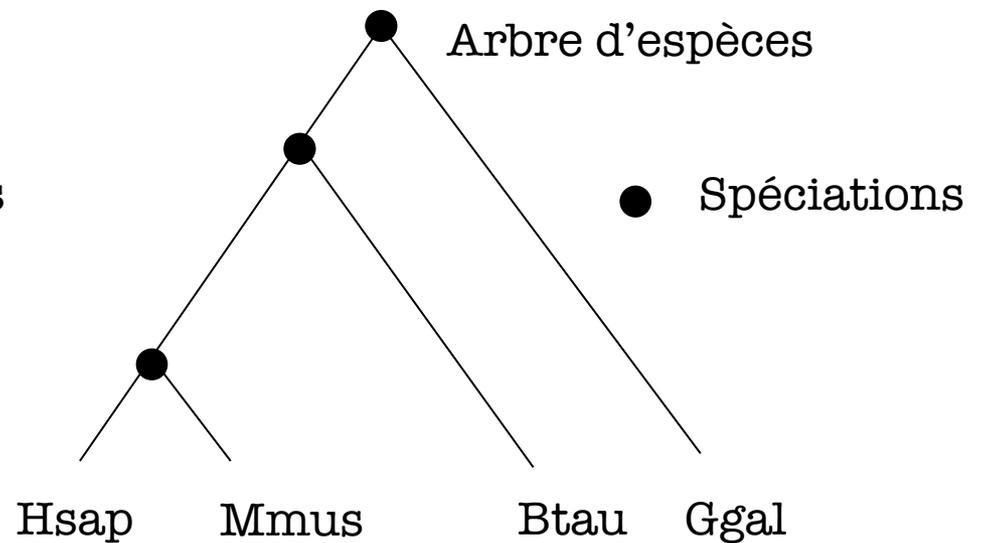


Expliquent la différence entre les arbres de gènes et les arbres d'espèces

Arbres de gènes



Arbre d'espèces



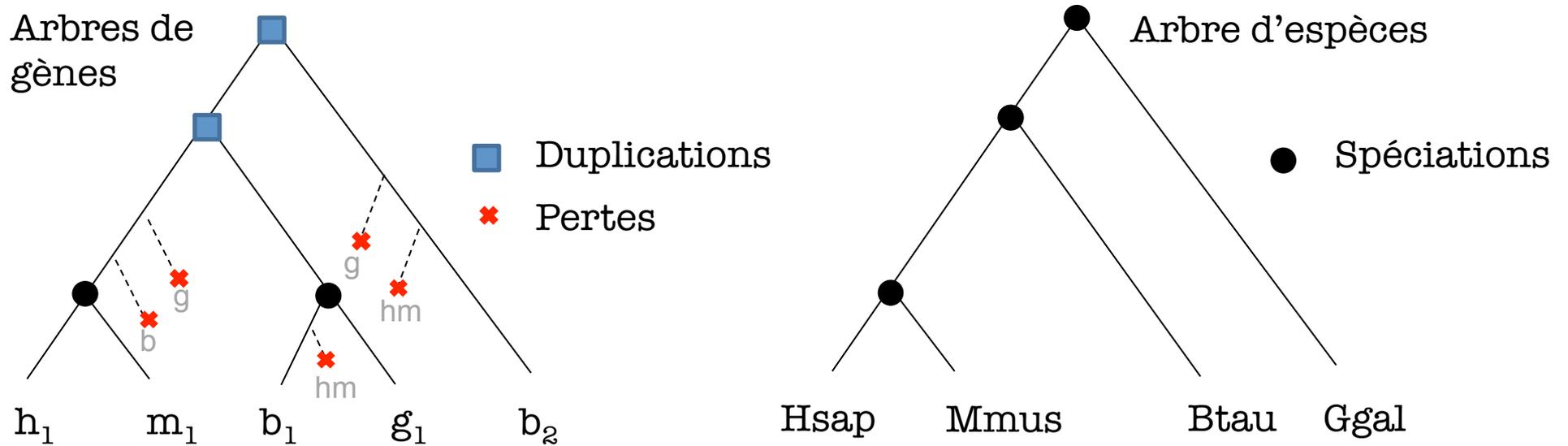
Ensemble de gènes homologues
Issus d'un même gène ancestral

Évolution des gènes

□ Considère les changements au niveau des gènes

- Gain de gènes
- Perte de gènes

Expliquent la différence entre les arbres de gènes et les arbres d'espèces



Ensemble de gènes homologues
Issus d'un même gène ancestral

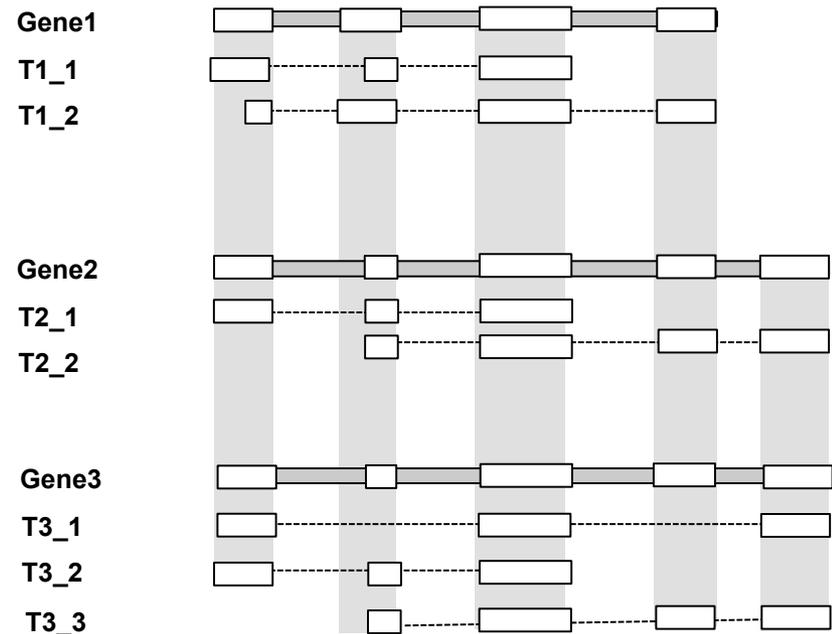
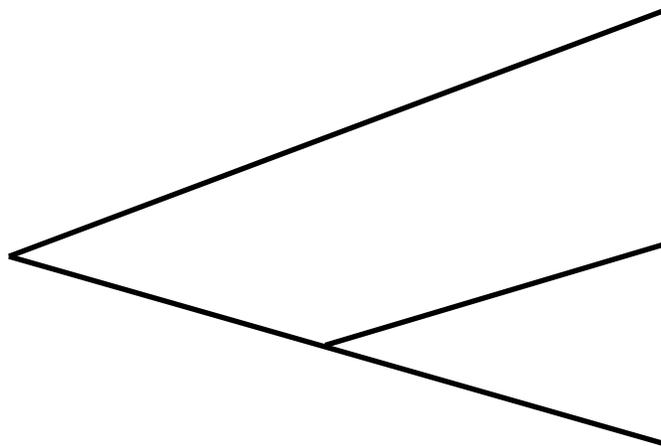
Évolution des transcrits

❑ Considère les changements de la structure des gènes et des ensembles d'ARNs

- Modification de la structure des gènes
- Modifications des ensembles d'ARNs



Expliquent la différence entre les arbres de transcrits et les arbres de gènes

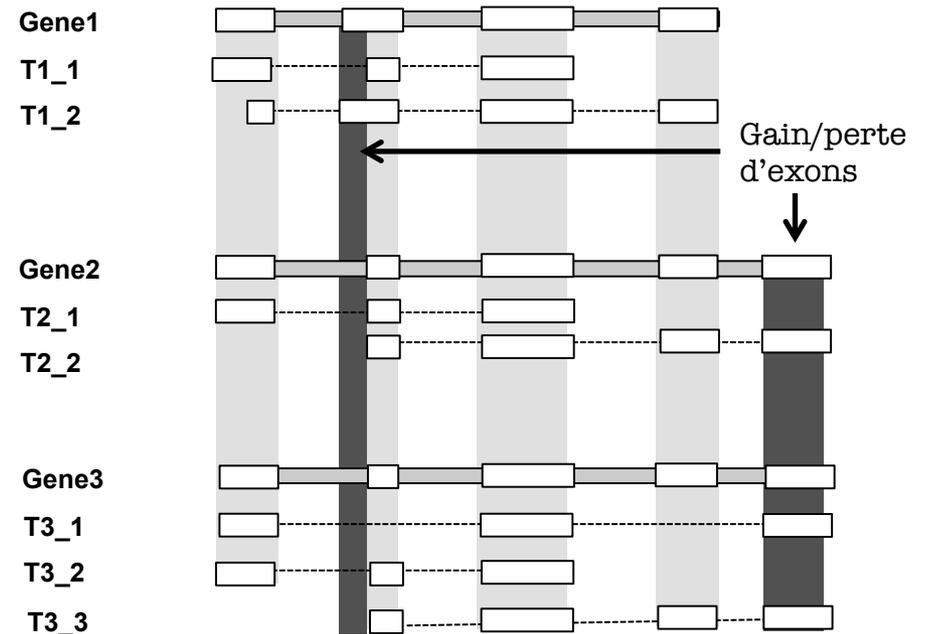
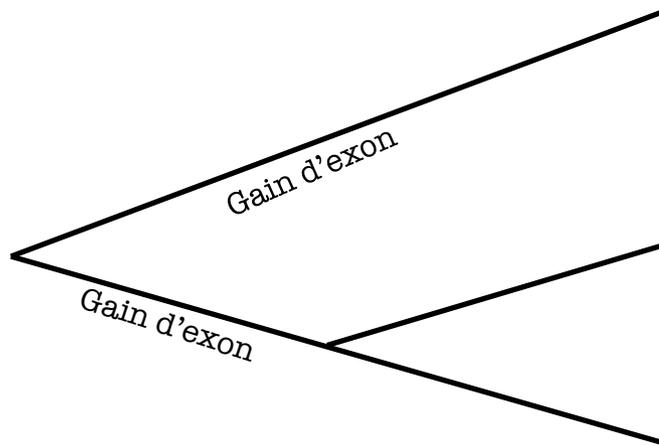


Évolution des transcrits

□ Considère les changements de la structure des gènes et des ensembles d'ARNs

- **Modification de la structure des gènes**
- Modifications des ensembles d'ARNs

Expliquent la différence entre les arbres de transcrits et les arbres de gènes

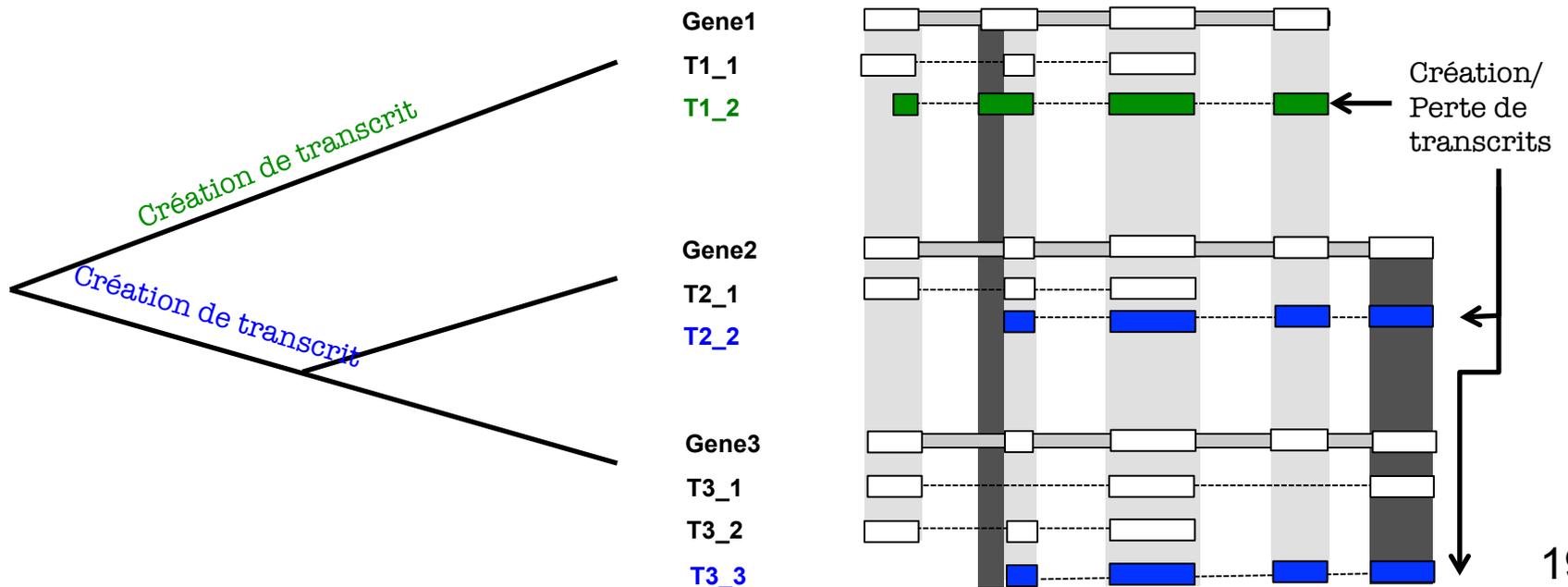


Évolution des transcrits

❑ Considère les changements de la structure des gènes et des ensembles d'ARNs

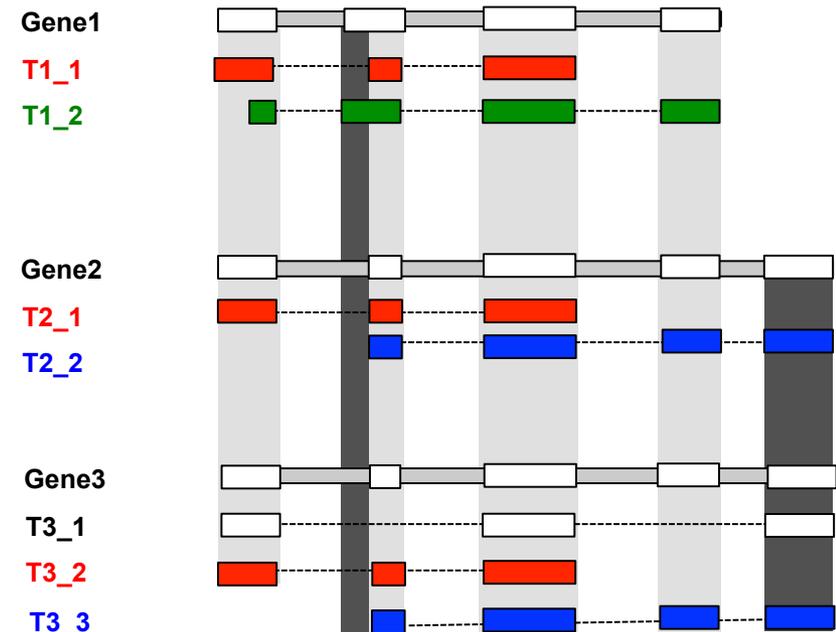
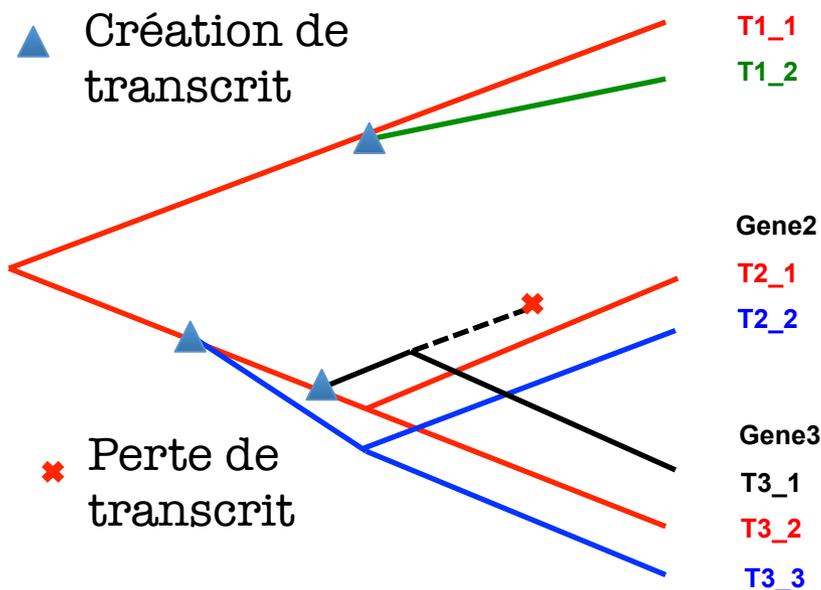
- Modification de la structure des gènes
- **Modifications des ensembles d'ARNs**

Expliquent la différence entre les arbres de transcrits et les arbres de gènes



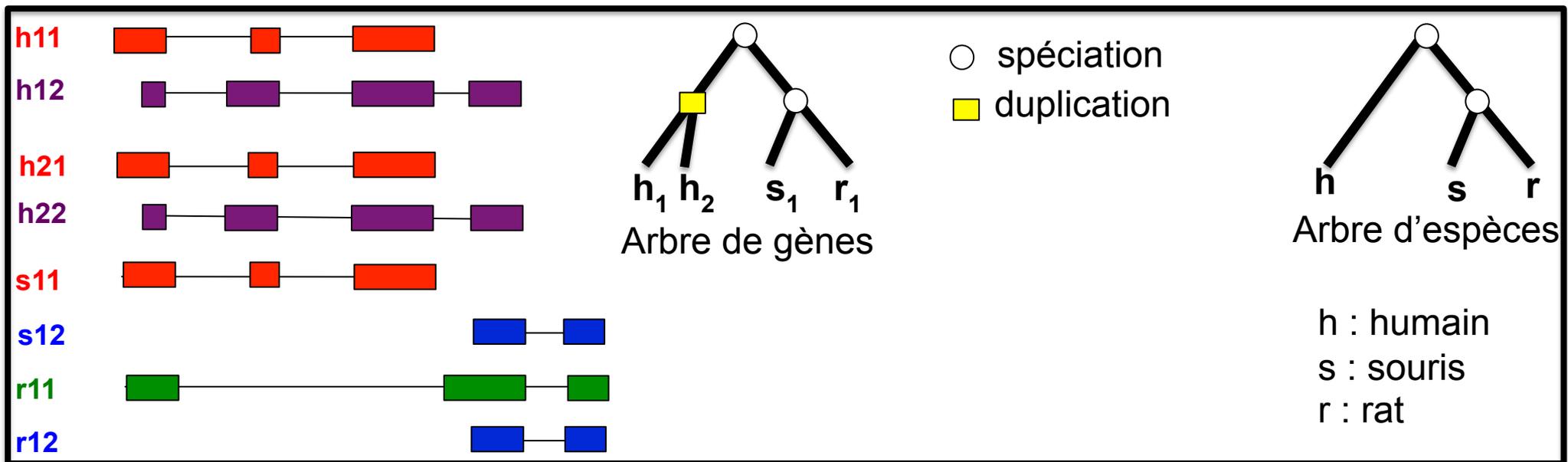
Modèle d'évolution des transcrits

- Considère les ensembles de transcrits conservés et les changements au niveau des ensembles de transcrits
 - Création de nouveaux transcrits
 - Perte de transcrits



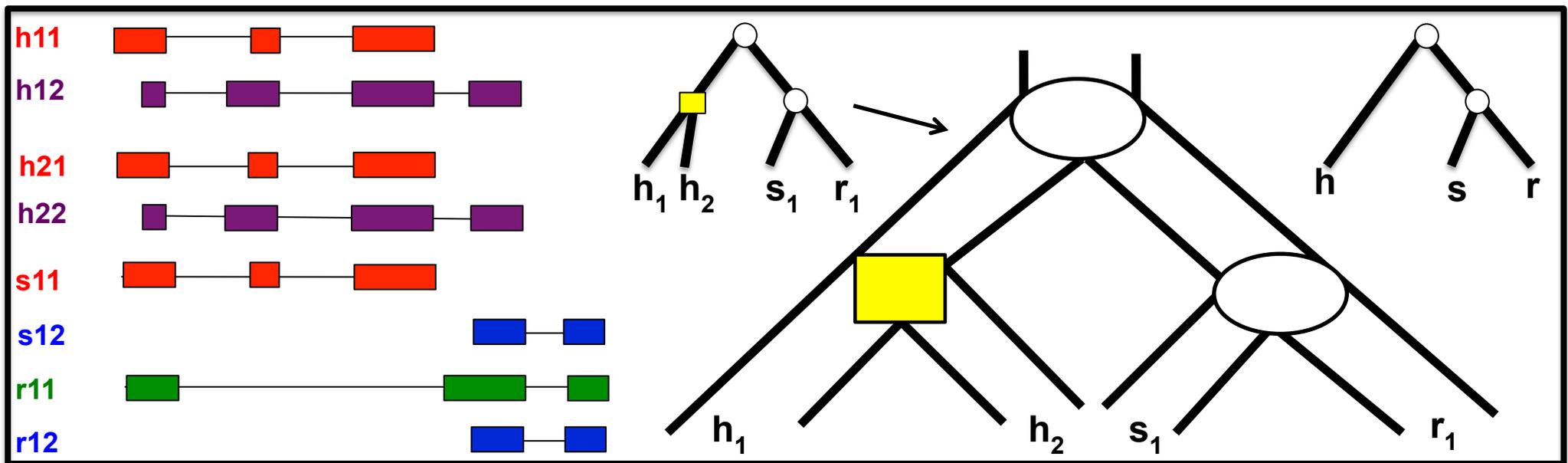
Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces



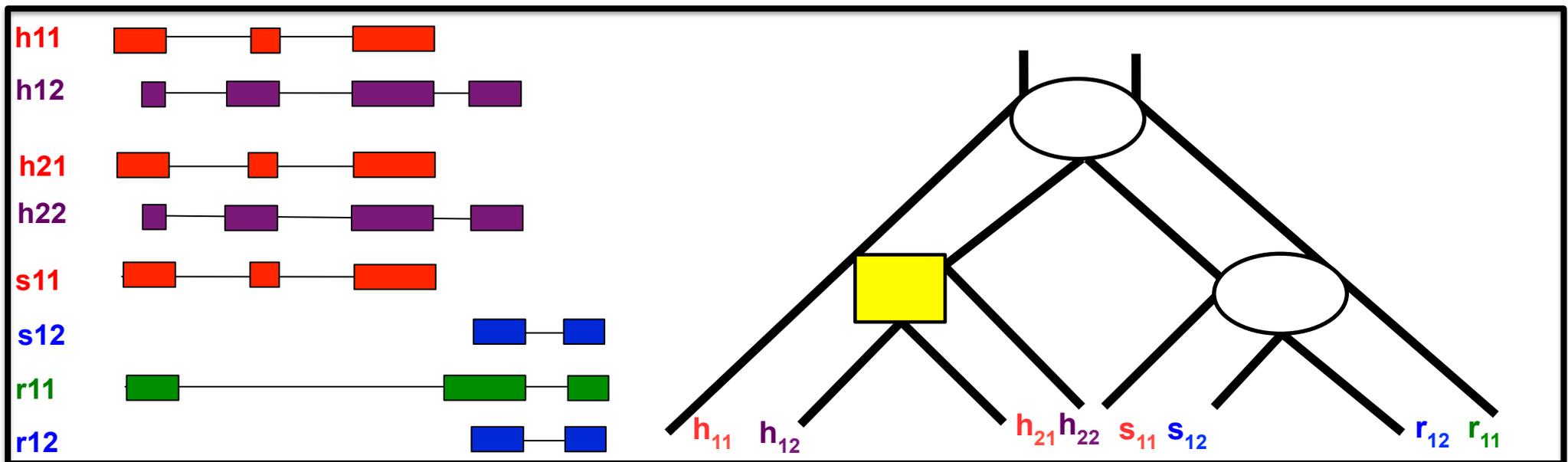
Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces



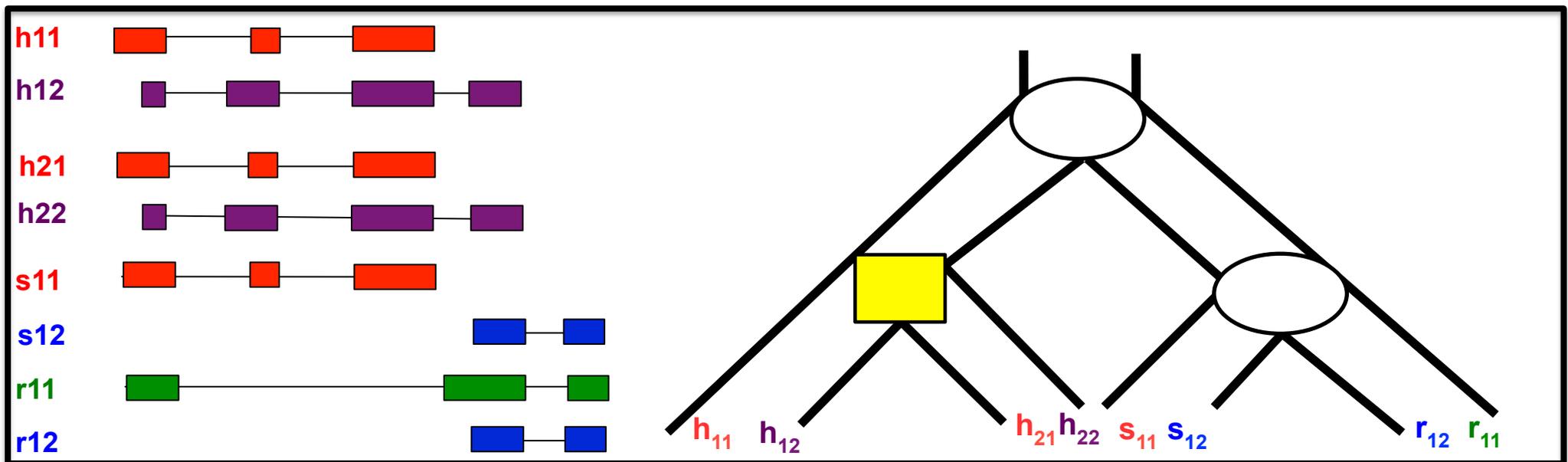
Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces



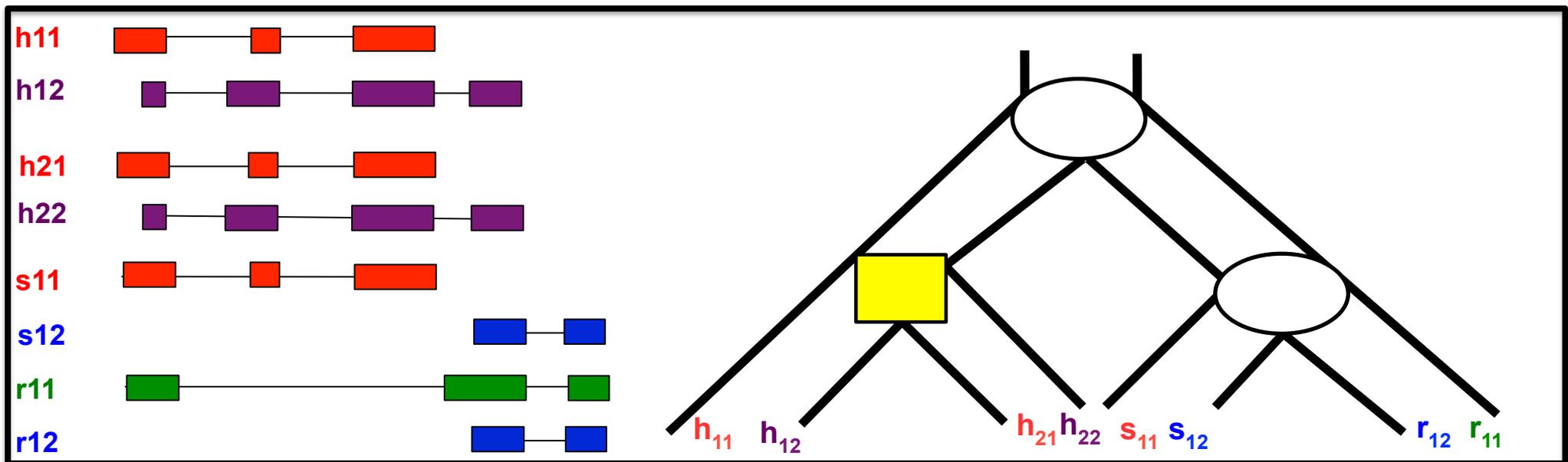
Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?



Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

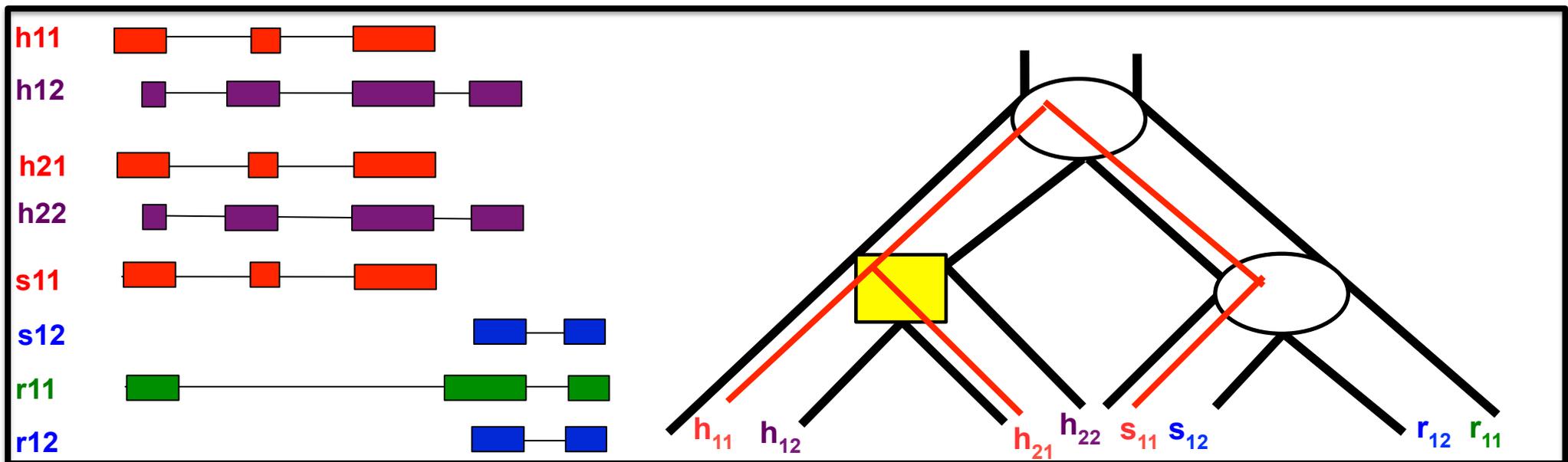


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie (caractères communs ne provenant pas d'un ancêtre commun)

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

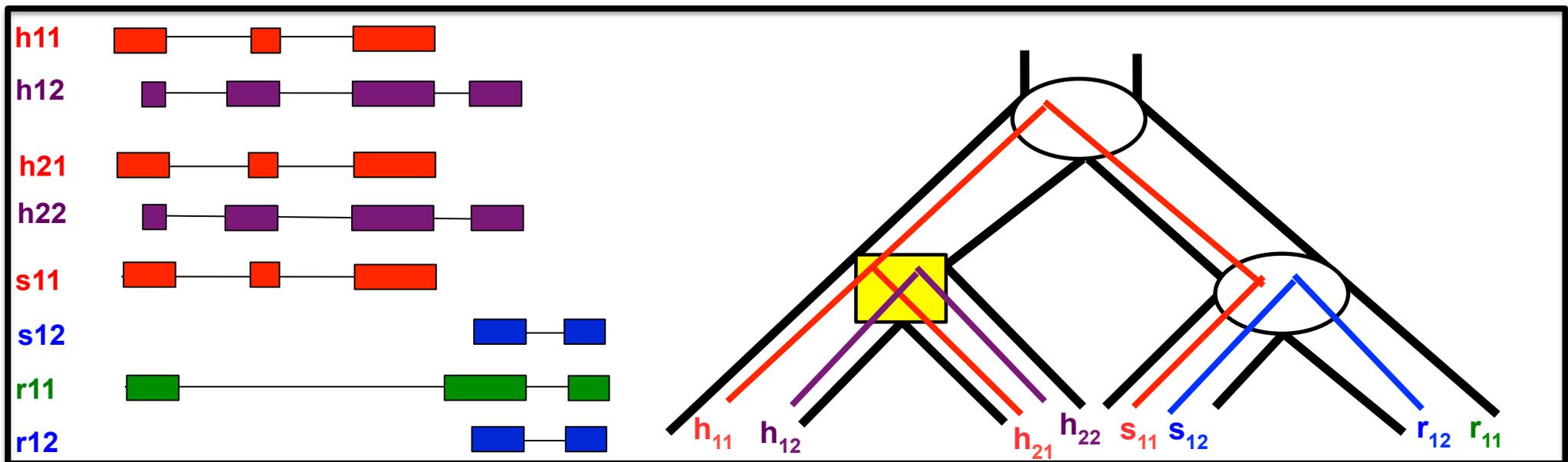


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

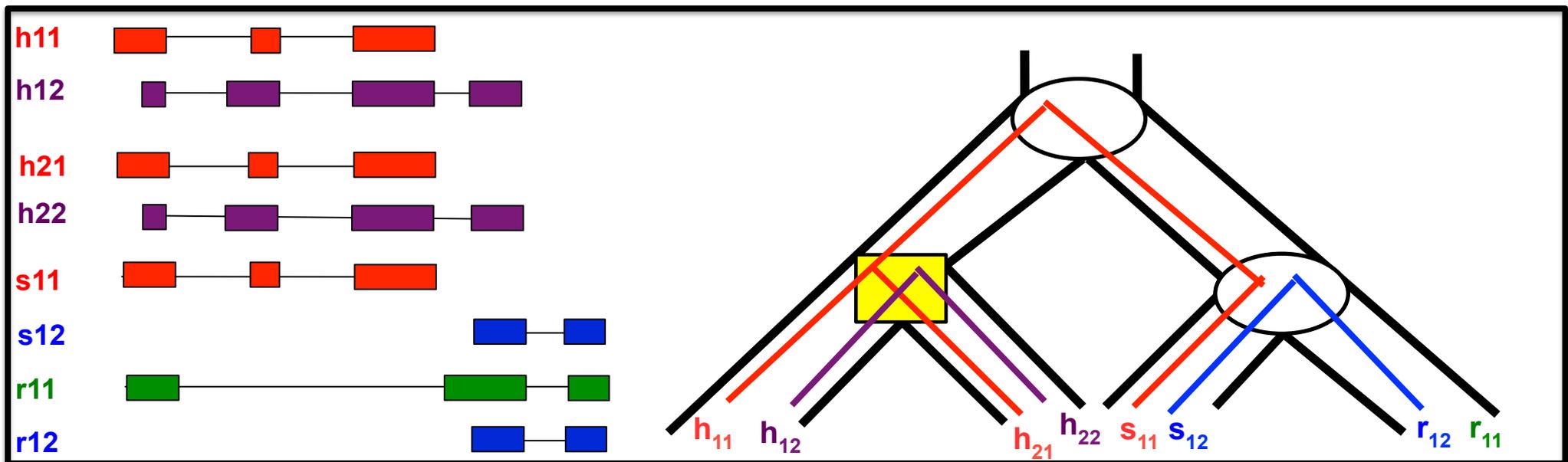


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

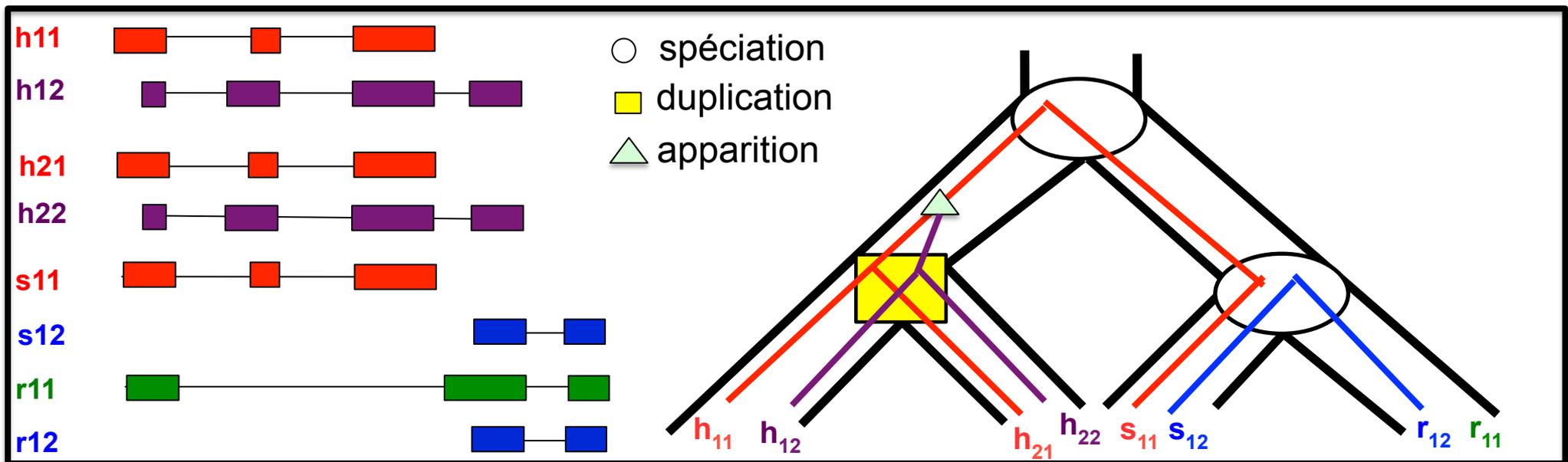


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie
- ◆ Les différences structurelles entre les transcrits sont dues à des événements de transcription et d'épissage alternatifs
 - minimisation de la distance évolutive

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

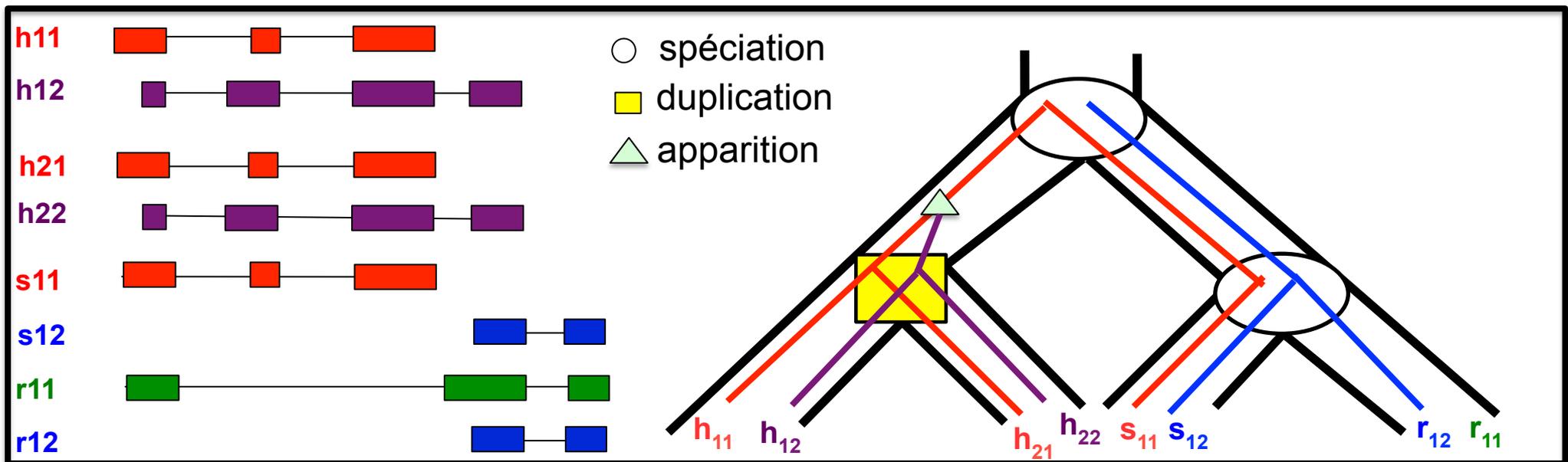


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie
- ◆ Les différences structurelles entre les transcrits sont dues à des événements de transcription et d'épissage alternatifs
 - minimisation de la distance évolutive

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

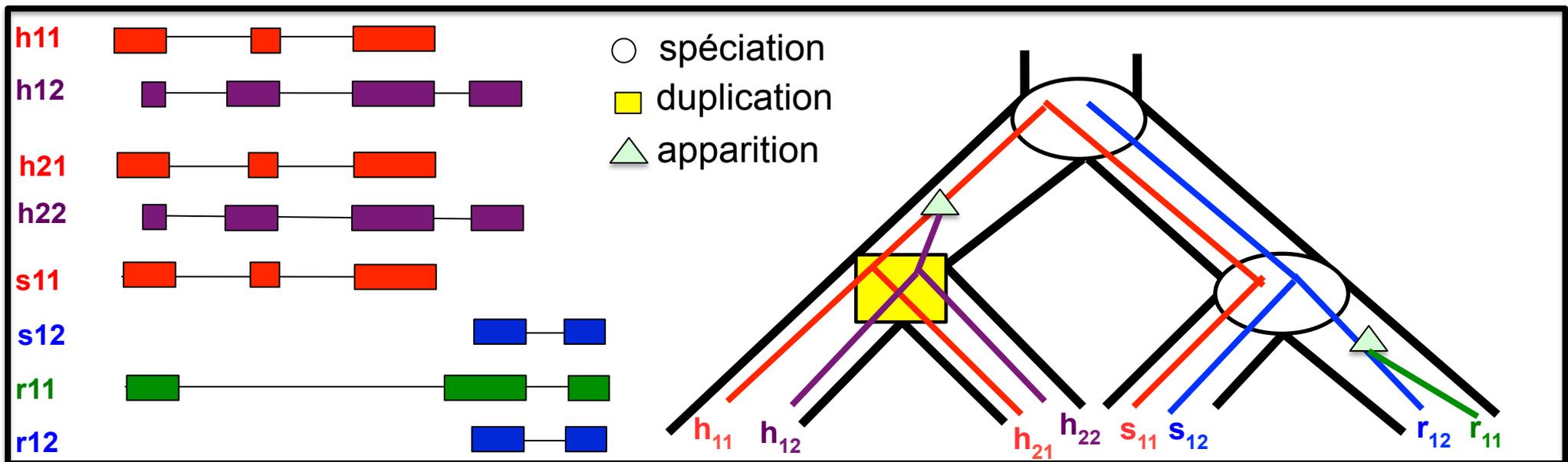


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie
- ◆ Les différences structurelles entre les transcrits sont dues à des événements de transcription et d'épissage alternatifs
 - minimisation de la distance évolutive

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

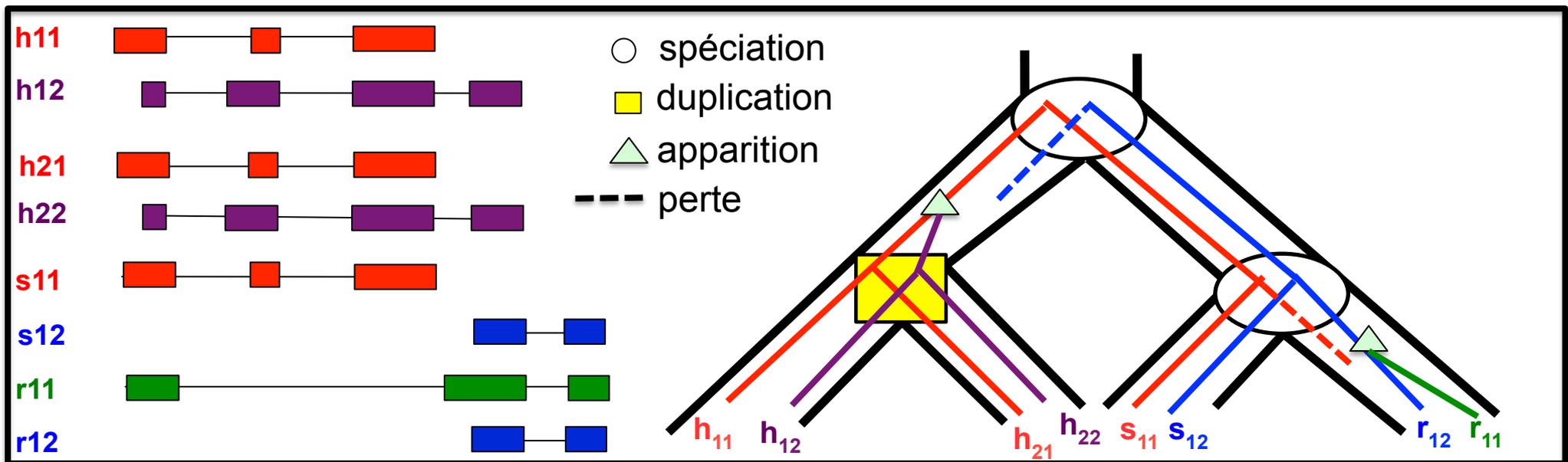


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie
- ◆ Les différences structurelles entre les transcrits sont dues à des événements de transcription et d'épissage alternatifs
 - minimisation de la distance évolutive

Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces
 - Quand est apparu un transcrit ? pourquoi (épissage alternatif, ...) ?
 - Quand a été perdu un transcrit ? pourquoi (mutation dans site d'épissage, ...) ?

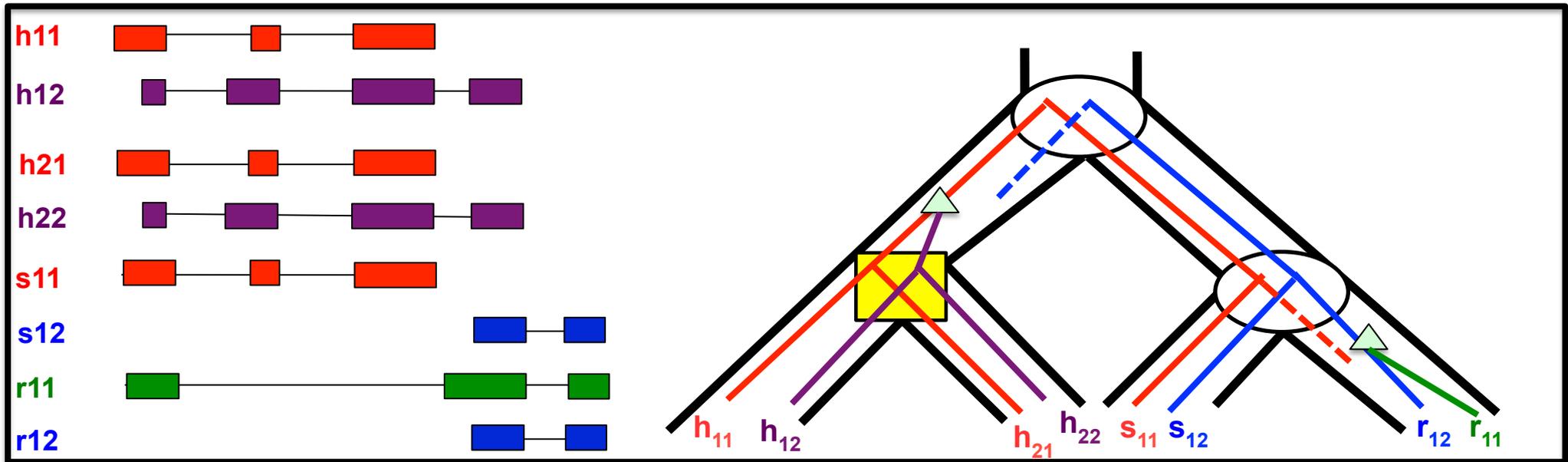


□ Modèle d'évolution sous-jacent

- ◆ Les transcrits communs observés descendent d'un même transcrit ancestral
 - minimisation de l'homoplasie
- ◆ Les différences structurelles entre les transcrits sont dues à des événements de transcription et d'épissage alternatifs
 - minimisation de la distance évolutive

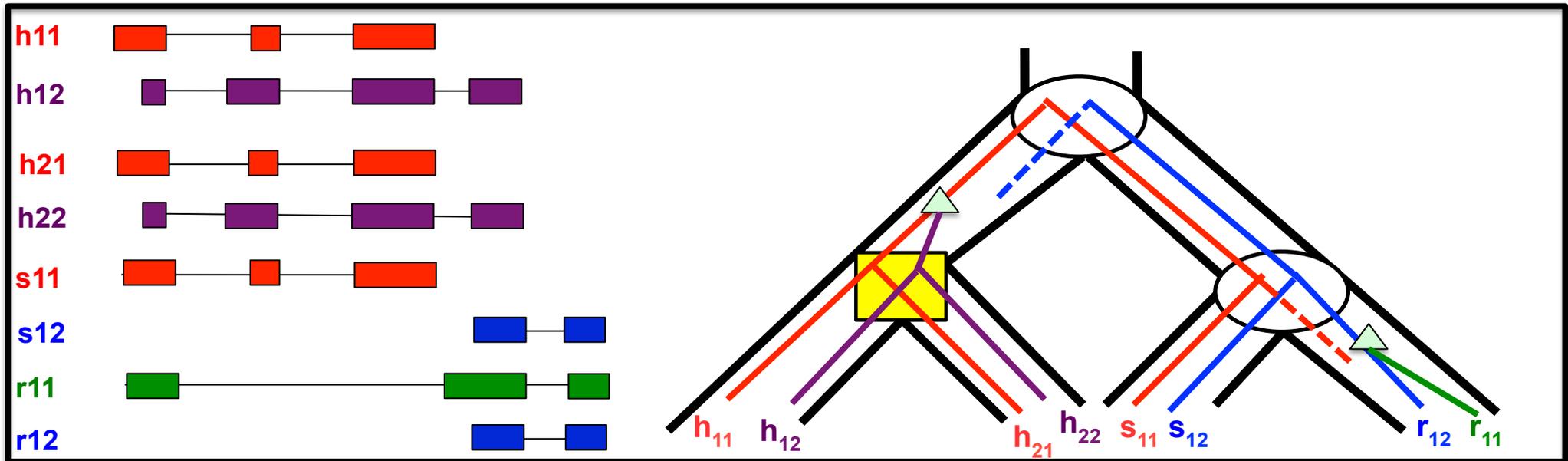
Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces



Évolution de la structure des gènes eucaryotes

- Inférer l'évolution d'ensembles de transcrits au sein d'une famille à la lumière de l'évolution des gènes et des espèces



□ Plusieurs étapes

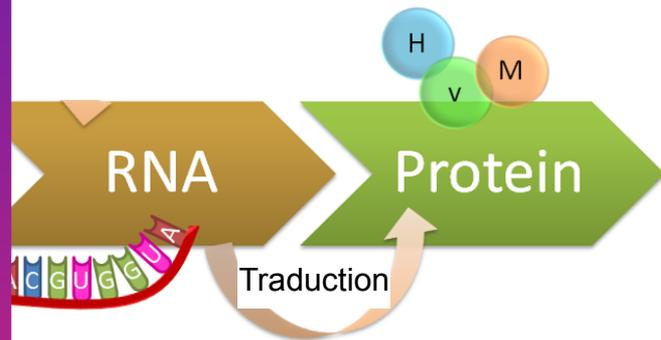
- ◆ Alignement des séquences de transcrits et gènes
- ◆ Identification des groupes de transcrits conservés
- ◆ Estimation de l'arbre des transcrits
- ◆ Réconciliation avec l'arbre des gènes et l'arbre des espèces
- ◆ Inférence d'un scénario d'épissage alternatif optimal sur les branches de l'arbre

Alignement des transcrits et gènes

- FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)
 - ◆ Alignement de séquences codantes **en tenant compte des longueurs de décalage de phase de traduction** pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Alignement des transcrits et gènes

- FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)
 - ◆ Alignement de séquences codantes **en tenant compte des longueurs de décalage de phase de traduction** pour l'estimation d'arbres de transcrits basée sur la distance évolutive



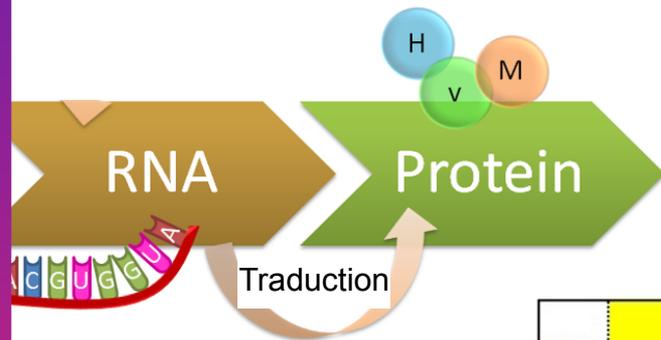
Codon(3 nucléotides (nt)) → Acide aminé (aa)

nt	AUG	ACC	GAA	UCC	AAG	CAG	CCC	UGG	CAU	AAG	UGG	GAG	UAA
aa	M	T	E	S	K	Q	P	W	H	K	W	E	*

Alignement des transcrits et gènes

□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

- ◆ Alignement de séquences codantes **en tenant compte des longueurs de décalage de phase de traduction** pour l'estimation d'arbres de transcrits basée sur la distance évolutive



Codon (3 nucléotides (nt)) → Acide aminé (aa)
 nt AUG**ACCGAA**U**CCAAG**C**AGCCC**U**GGCAU**A**AGUGGGAG**UAA
 aa M **T** E **S** K **Q** P **W** H **K** W **E** *

	U	C	A	G	
U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U
	UUC } Ser	UCC } Ser	UAC } Tyr	UGC } Cys	C
	UUA }	UCA }	UAA }	UGA }	A
	UUG }	UCG }	UAG }	UGG } Trp	G
C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U
	CUC } Leu	CCC } Pro	CAC } His	CGC } Arg	C
	CUA }	CCA } Pro	CAA } Gln	CGA } Arg	A
	CUG }	CCG } Pro	CAG } Gln	CGG } Arg	G
A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U
	AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser	C
	AUA }	ACA } Thr	AAA } Lys	AGA } Arg	A
	AUG } Met	ACG } Thr	AAG } Lys	AGG } Arg	G
G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U
	GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly	C
	GUA }	GCA } Ala	GAA } Glu	GGA } Gly	A
	GUG }	GCG } Ala	GAG } Glu	GGG } Gly	G

- ◆ 64 codons
- ◆ 20 acides aminés

Alignement des transcrits et gènes

- FsePSA (S. Jammali et al., *Algorithms for Molecular Biology* 2017)
 - ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

```
S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
***** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** ** **
```

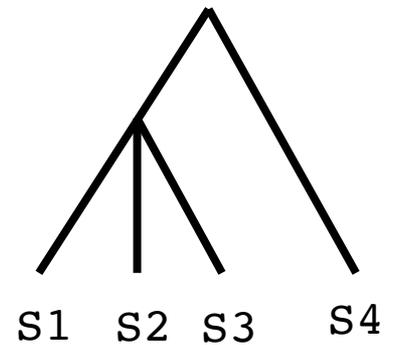
Alignement des transcrits et gènes

□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

- ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

```
S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
*****
```



Alignement des transcrits et gènes

□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

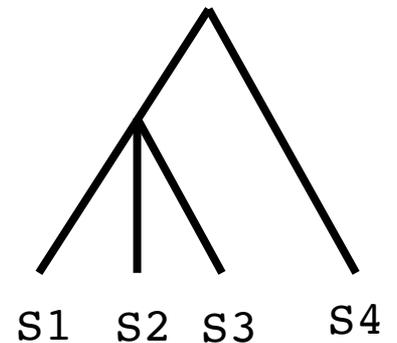
- ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

```

S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
*****

```



Acide aminé (aa)

S1	M	T	E	S	K	Q	P	W	H	K	W	E	*
S4	M	T	E	S	N	Q	P	C	H	K	C	E	*
	*	*	*	*	*	*	*	*	*	*	*	*	*
S2	M	T	E	S	K	Q	P	W	H	N	G	S	*
	*	*	*	*	:	*	*	:	*				
S3	M	T	E	S	N	S	P	G	I	S	G	S	*
	*	*	*	*									



Alignement des transcrits et gènes

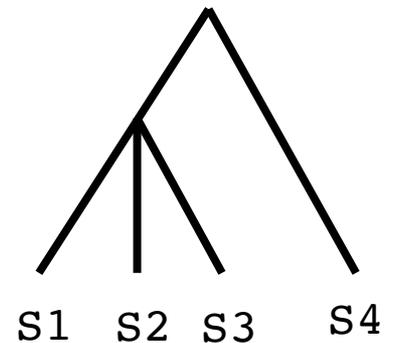
□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

- ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

```

S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
*****
  
```



Acide aminé (aa)

S1	M	T	E	S	K	Q	P	W	H	K	W	E	*
S4	M	T	E	S	N	Q	P	C	H	K	C	E	*
	*	*	*	*	*	*	*	*	*	*		*	
S2	M	T	E	S	K	Q	P	W	H	N	G	S	*
	*	*	*	*	:	*	*	:	*				
S3	M	T	E	S	N	S	P	G	I	S	G	S	*
	*	*	*	*									



Décalage ←→

Alignement des transcrits et gènes

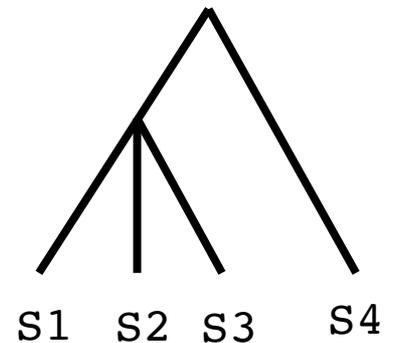
□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

- ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

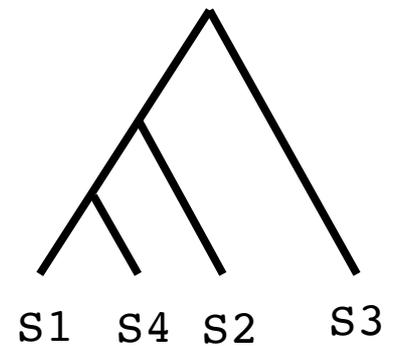
```

S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
*****
  
```



Acide aminé (aa)

S1	M	T	E	S	K	Q	P	W	H	K	W	E	*
S4	M	T	E	S	N	Q	P	C	H	K	C	E	*
	*	*	*	*	*	*	*	*	*	*	*	*	
S2	M	T	E	S	K	Q	P	W	H	N	G	S	*
	*	*	*	*	:	*	*	:	*	↔			
S3	M	T	E	S	N	S	P	G	I	S	G	S	*
	*	*	*	*	↔								



Décalage ↔

Alignement des transcrits et gènes

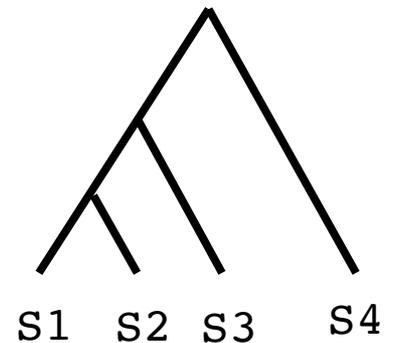
□ FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)

- ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction pour l'estimation d'arbres de transcrits basée sur la distance évolutive

Nucléotide (nt)

```

S1 AUGACCGAAUCCAAGCAGCCCUGGCAUAAGUGGGAGUAA-
S2 AUGACCGAAUCCAAGCAGCCCUGGCAUAA-UGGGAGUUAG
S3 AUGACCGAAUCCAA-CAGCCCUGGCAUAAGUGGGAGUUAG
*****
S4 AUGACUGAGUCAAAACCAACCGUGCCACAAAUGCGAAUAA-
*****
  
```



Acide aminé (aa)

S1	M	T	E	S	K	Q	P	W	H	K	W	E	*
S4	M	T	E	S	N	Q	P	C	H	K	C	E	*
	*	*	*	*	*	*	*	*	*	*	*	*	
S2	M	T	E	S	K	Q	P	W	H	N	G	S	*
	*	*	*	*	:	*	*	:	*	↔			
S3	M	T	E	S	N	S	P	G	I	S	G	S	*
	*	*	*	*	↔							↔	

Décalage ↔

Alignement des transcrits et gènes

- FsePSA (S. Jammali et al., Algorithms for Molecular Biology 2017)
 - ◆ Alignement de séquences codantes en tenant compte des longueurs de décalage de phase de traduction
 - ◆ Algorithme de programmation dynamique multi-dimensionnel
- Score basé sur une partition des codons en 4 ensembles dépendant de leur alignement
 - ◆ En phase de traduction conservée
 - ◆ En phase de traduction décalée
 - ◆ Inséré/Supprimé
 - ◆ En début de décalage

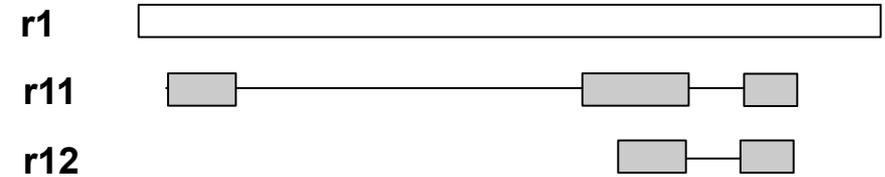
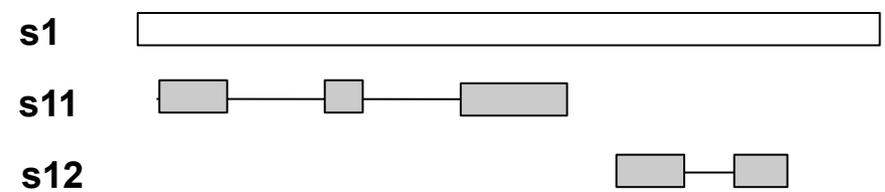
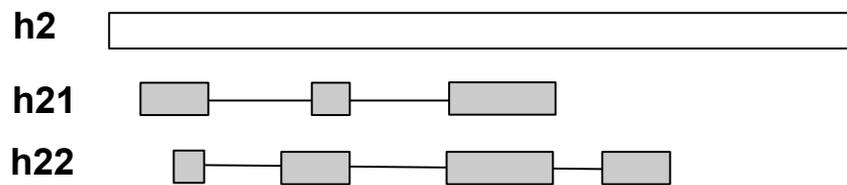
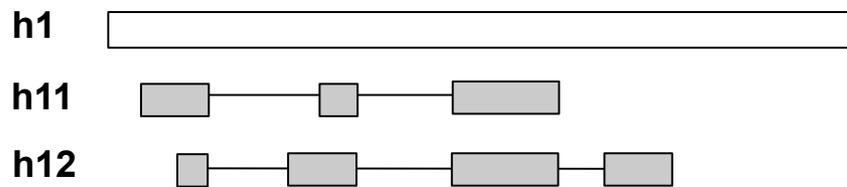
M	T	E	S	K		Q	P	W	P		D		Q	R	*	
AUG	ACC	GAA	UCC	AAG	--	CAG	CCC	UGG	CC	CAG	---	AU	---	CAA	CG	-UUGA
AUG	---	GAG	UCG	AAG	AUC	AGC	--	UGG	-	CAG	GC	CAU	UGG	CAA	UG	ACUGA
M		E	S	K	I	S		W		Q	A	I	G	N	D	*

- Score d'un alignement = somme des scores assignés aux alignements de codons
- Algorithme en temps $O(nm)$ pour le calcul d'un alignement de score optimal

Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

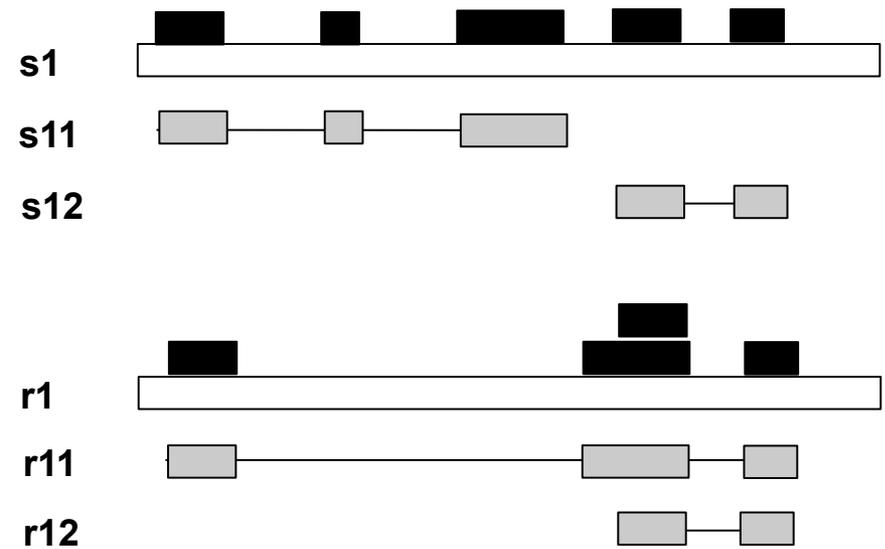
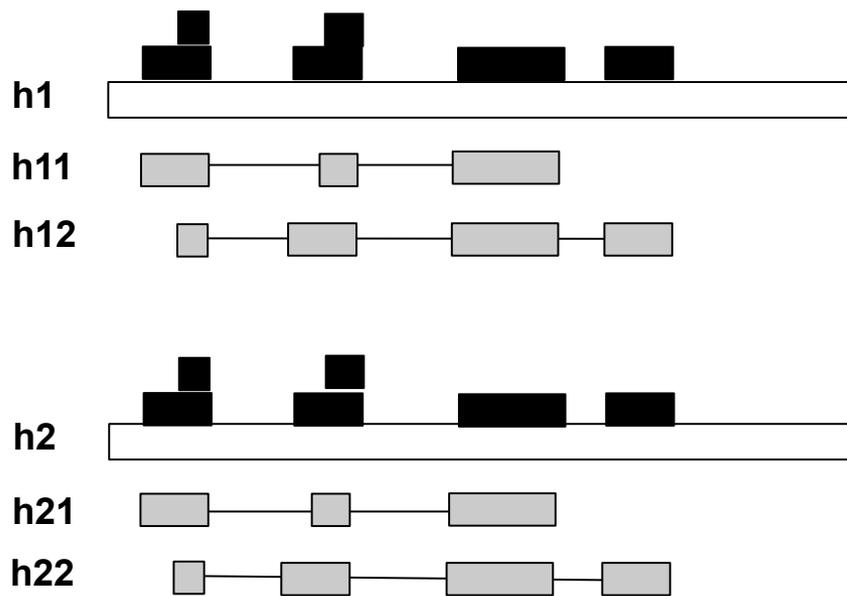
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

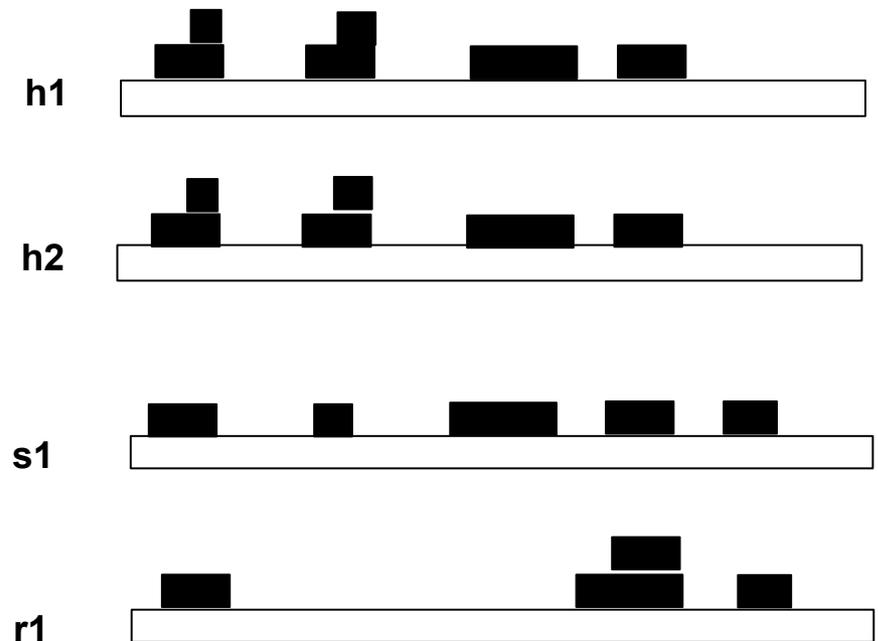
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

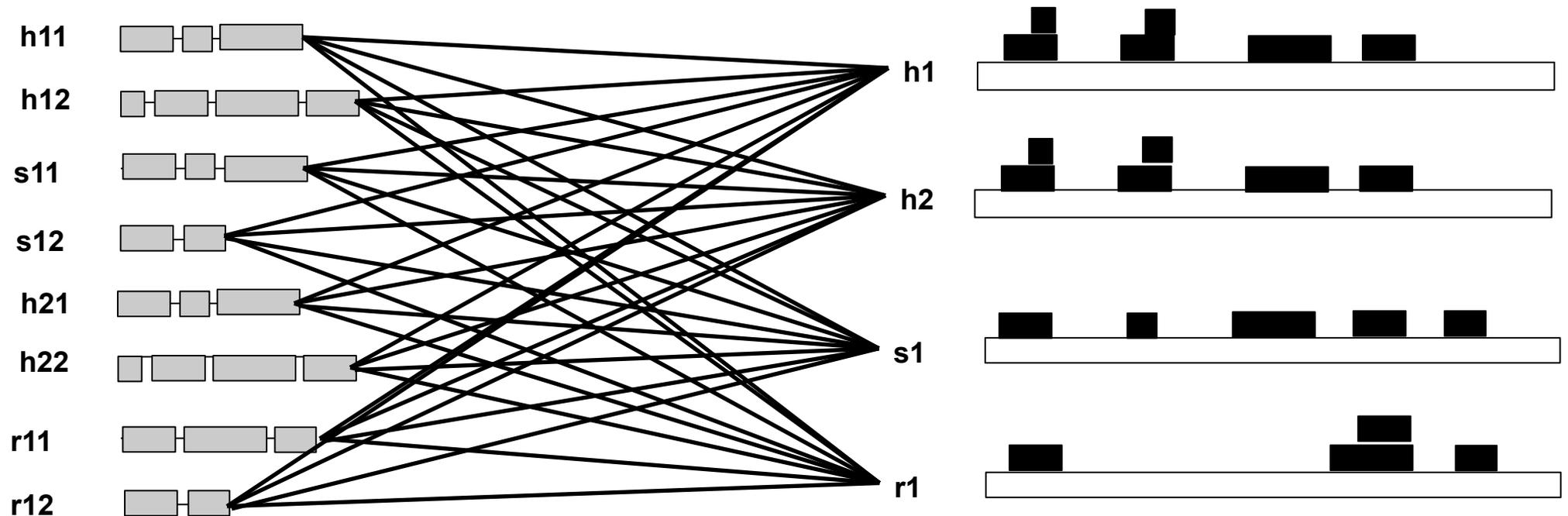
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

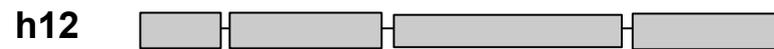
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

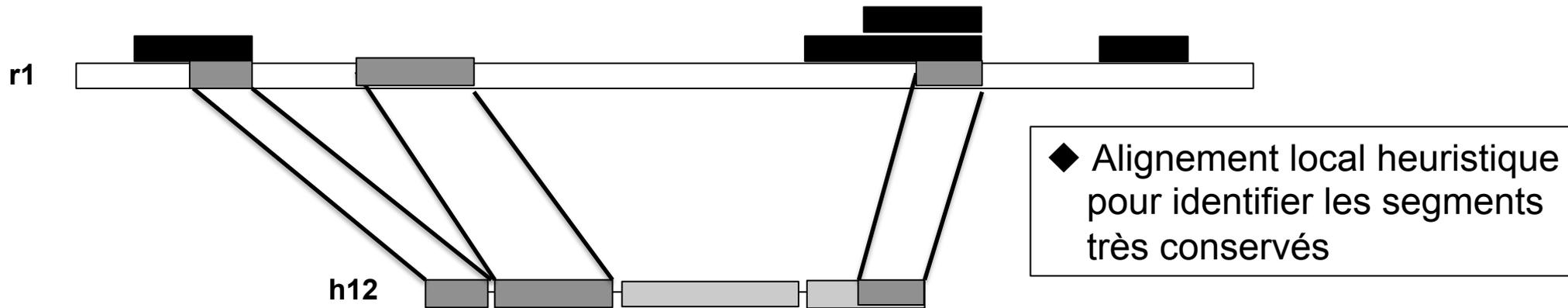
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

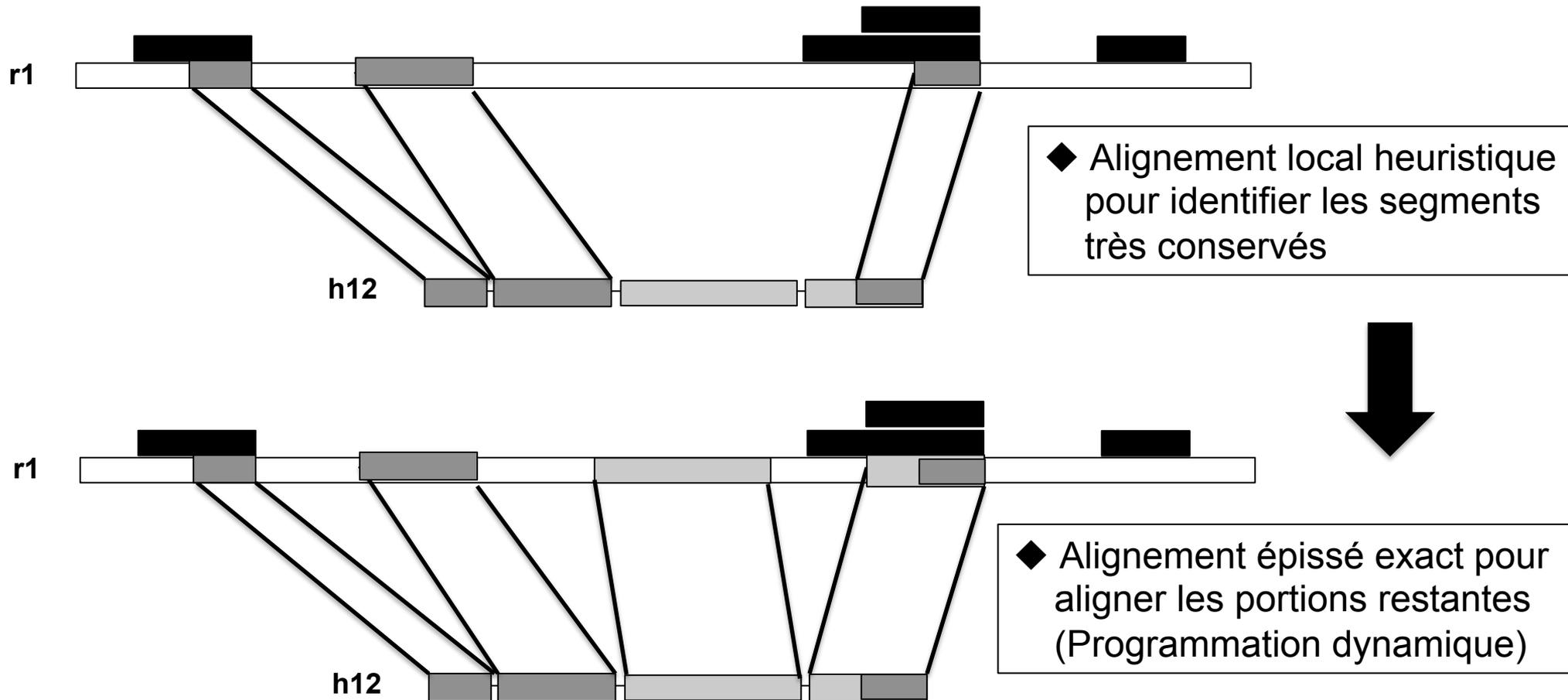
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Alignement des transcrits et gènes

□ SpliceFamAlign (J.-D. Aguilar et al., arXiv 2017)

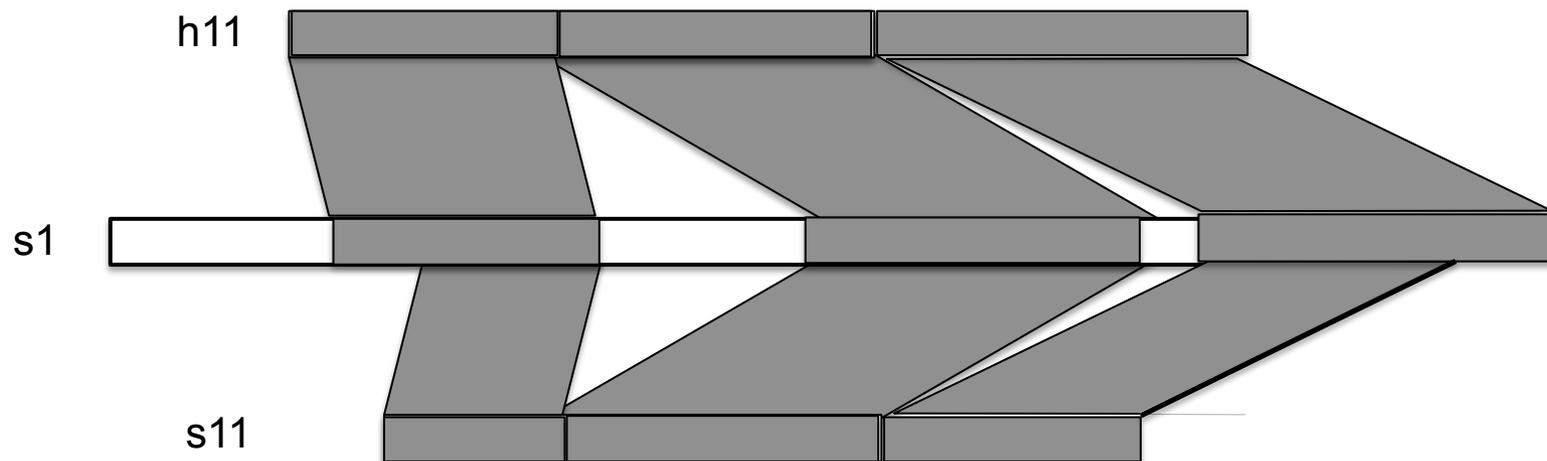
- ◆ Alignement épissé d'une séquence codante sur une séquence de gène **en tenant compte de la structure exon-intron des séquences** pour l'identification de groupes de transcrits conservés



Identification de transcrits conservés

□ SpliceFamAlign-Ortholog (J.-D. Aguilar et al., arXiv 2017)

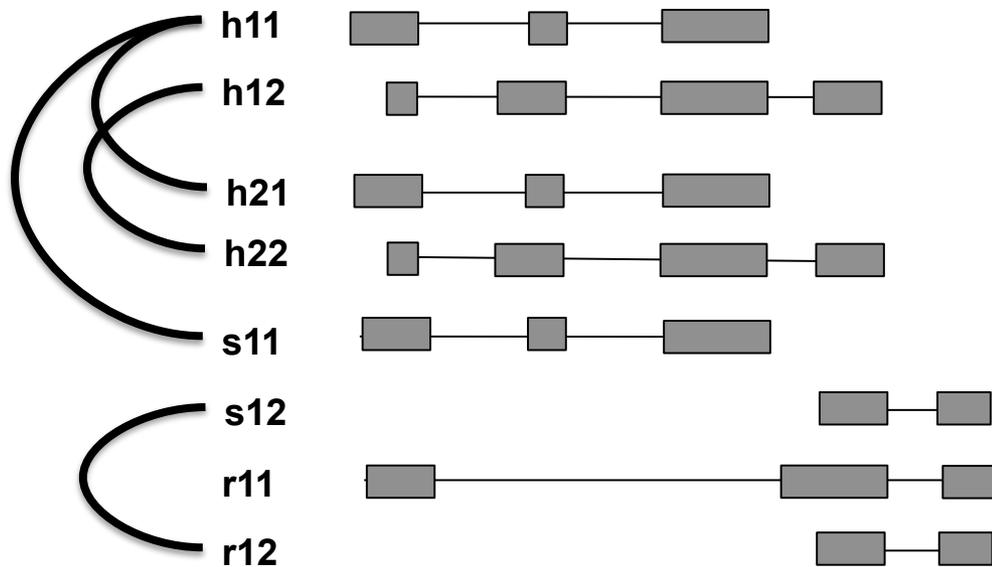
- ◆ **Paire de séquences conservées** : leurs alignements épissés contre un gène:
- préserve la structure en exon des deux séquences
 - induit une correspondance entre les exons des deux séquences



Identification de transcrits conservés

□ SpliceFamAlign-Ortholog (J.-D. Aguilar et al., arXiv 2017)

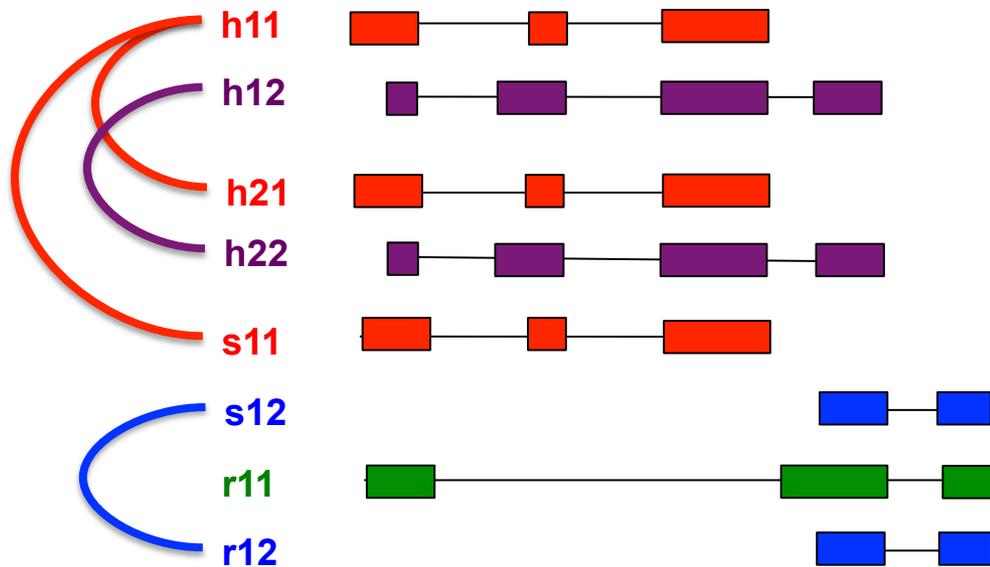
◆ **Groupe de transcrits conservés** : composantes connexes du graphe des transcrits



Identification de transcrits conservés

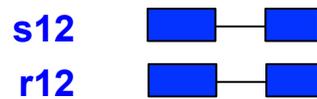
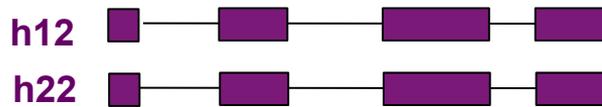
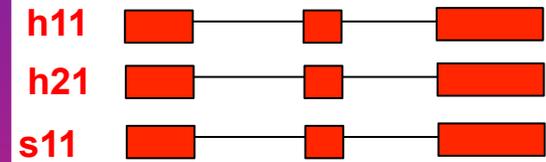
□ SpliceFamAlign-Ortholog (J.-D. Aguilar et al., arXiv 2017)

◆ **Groupe de transcrits conservés** : composantes connexes du graphe des transcrits



Construction d'un arbre des transcrits

- SuperProteinTree (E. Kuitche et al., J. of Bioinformatics and Comput. Biology 2017)
 - ◆ Estimation d'un arbre des transcrits à partir des groupes de transcrits conservés

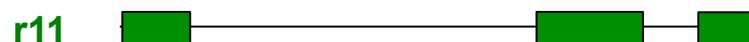
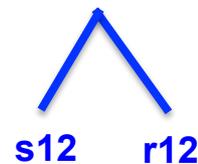
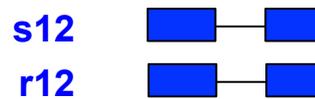
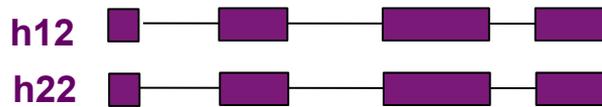
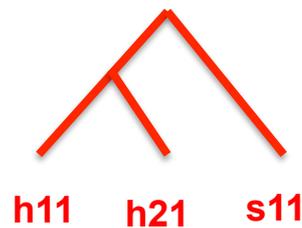
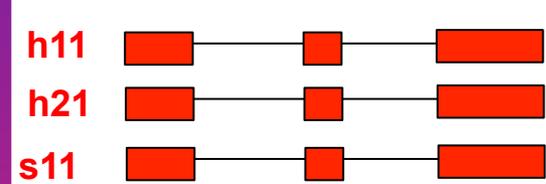


Construction d'un arbre des transcrits

- SuperProteinTree (E. Kuitche et al., J. of Bioinformatics and Comput. Biology 2017)
 - ◆ Estimation d'un arbre des transcrits à partir des groupes de transcrits conservés



Construire un arbre pour chaque groupe



r11

Construction d'un arbre des transcrits

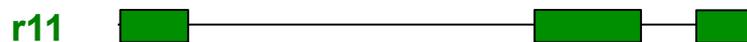
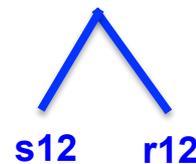
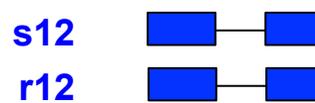
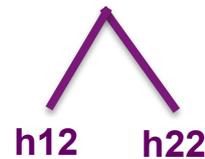
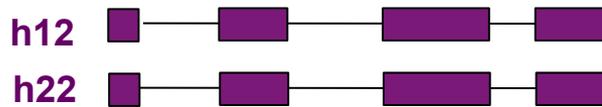
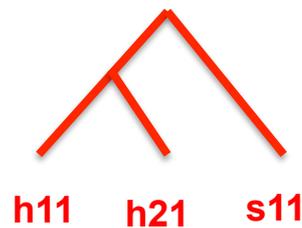
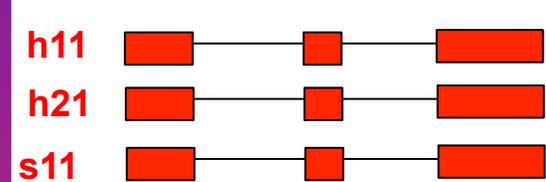
- SuperProteinTree (E. Kuitche et al., J. of Bioinformatics and Comput. Biology 2017)
 - ◆ Estimation d'un arbre des transcrits à partir des groupes de transcrits conservés



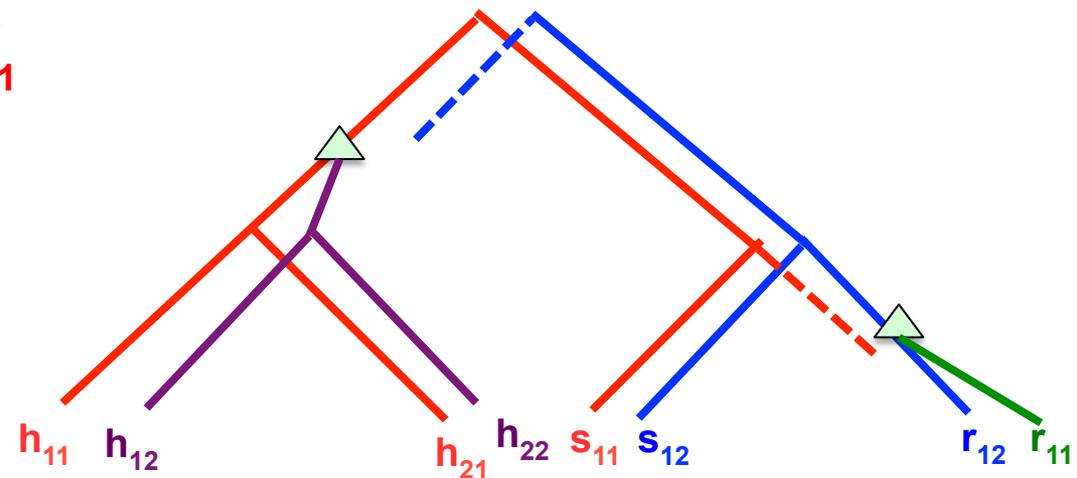
Construire un arbre pour chaque groupe



Regrouper les arbres en un super-arbre minimisant le nombre d'événements d'apparition, perte, et épissage alternatif



r11

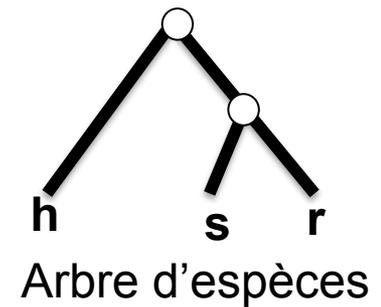
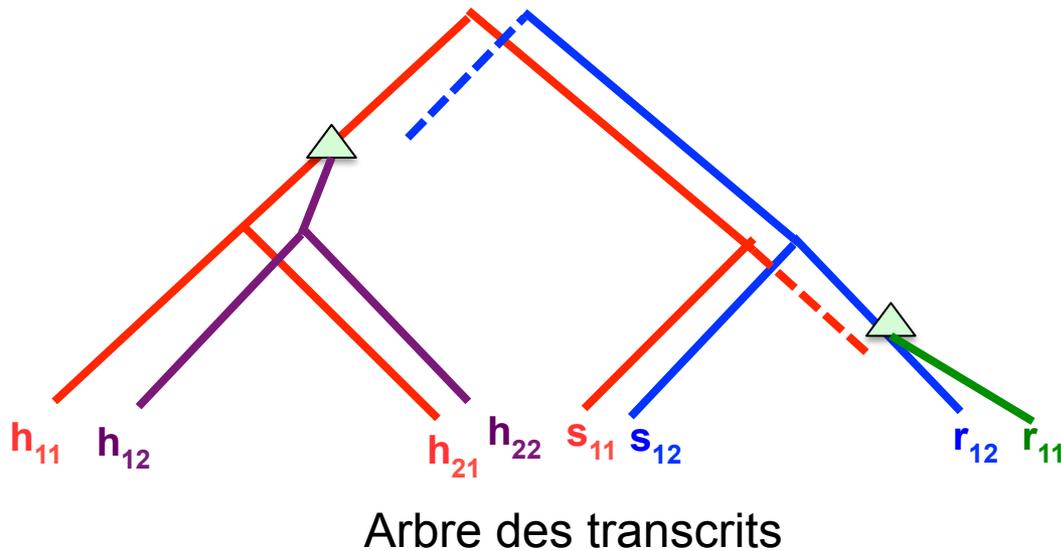


- ◆ Algorithme heuristique basé sur une approche constructive (classification hiérarchique)

Reconciliation avec arbre de gènes et arbre d'espèces

□ Protein2GeneTree (E. Kuitche et al., BICOB 2017)

- ◆ Estimation d'un arbre de gènes optimal, étant donné l'arbre des transcrits et l'arbre des espèces



Arbre des gènes ?

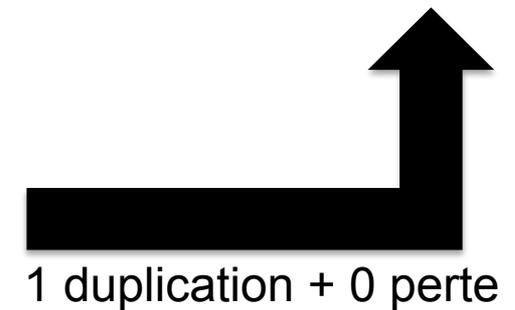
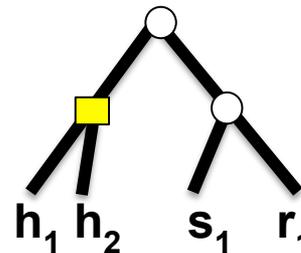
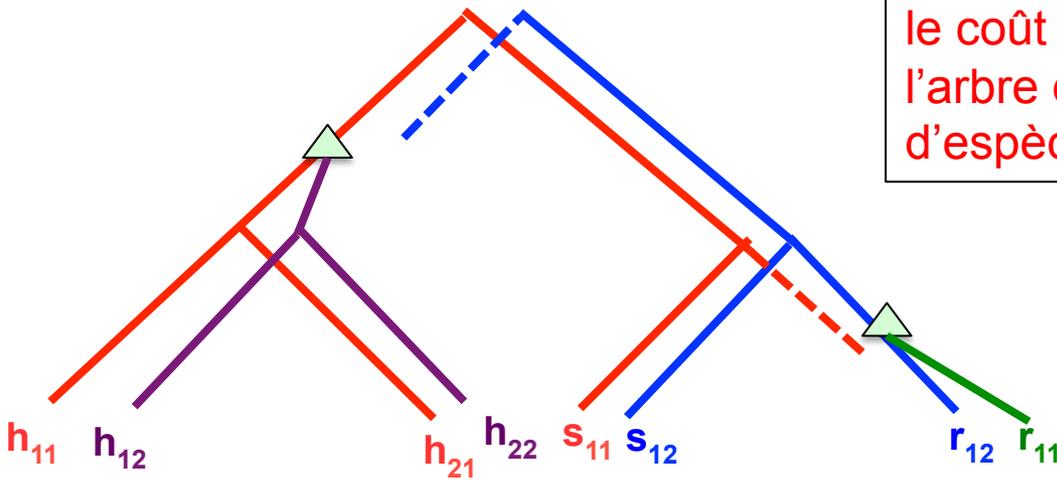
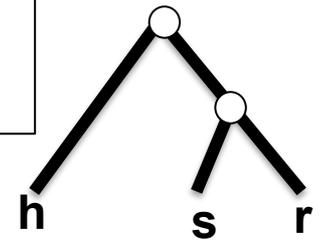
- ◆ Problème NP-dur

Reconciliation avec arbre de gènes et arbre d'espèces

□ Protein2GeneTree (E. Kuitche et al., BICOB 2017)

- ◆ Estimation d'un arbre de gènes optimal, étant donné l'arbre des transcrits et l'arbre des espèces

Trouver un arbre de gènes minimisant le coût de double réconciliation avec l'arbre des transcrits et l'arbre d'espèces



- ◆ Algorithme heuristique basé sur une approche de recherche locale à partir de l'arbres des transcrits

Conclusion et Perspectives

- Suite d'algorithmes pour la reconstruction de l'évolution de la structure de gènes
 - ◆ FsePSA
 - ◆ SpliceFamAlign
 - ◆ SpliceFamAlign-ortholog
 - ◆ SuperProteinTree
 - ◆ Protein2GeneTree

- Perspectives
 - ◆ Extension vers des alignements multiples des séquences codantes
 - ◆ Simulation de l'évolution d'ensembles de transcrits de gènes homologues
 - ◆ Modélisation de la structure des gènes pour la prédiction et la classification

<https://github.com/UdeS-CoBIUS/>

<http://compbio.dinf.usherbrooke.ca/>

