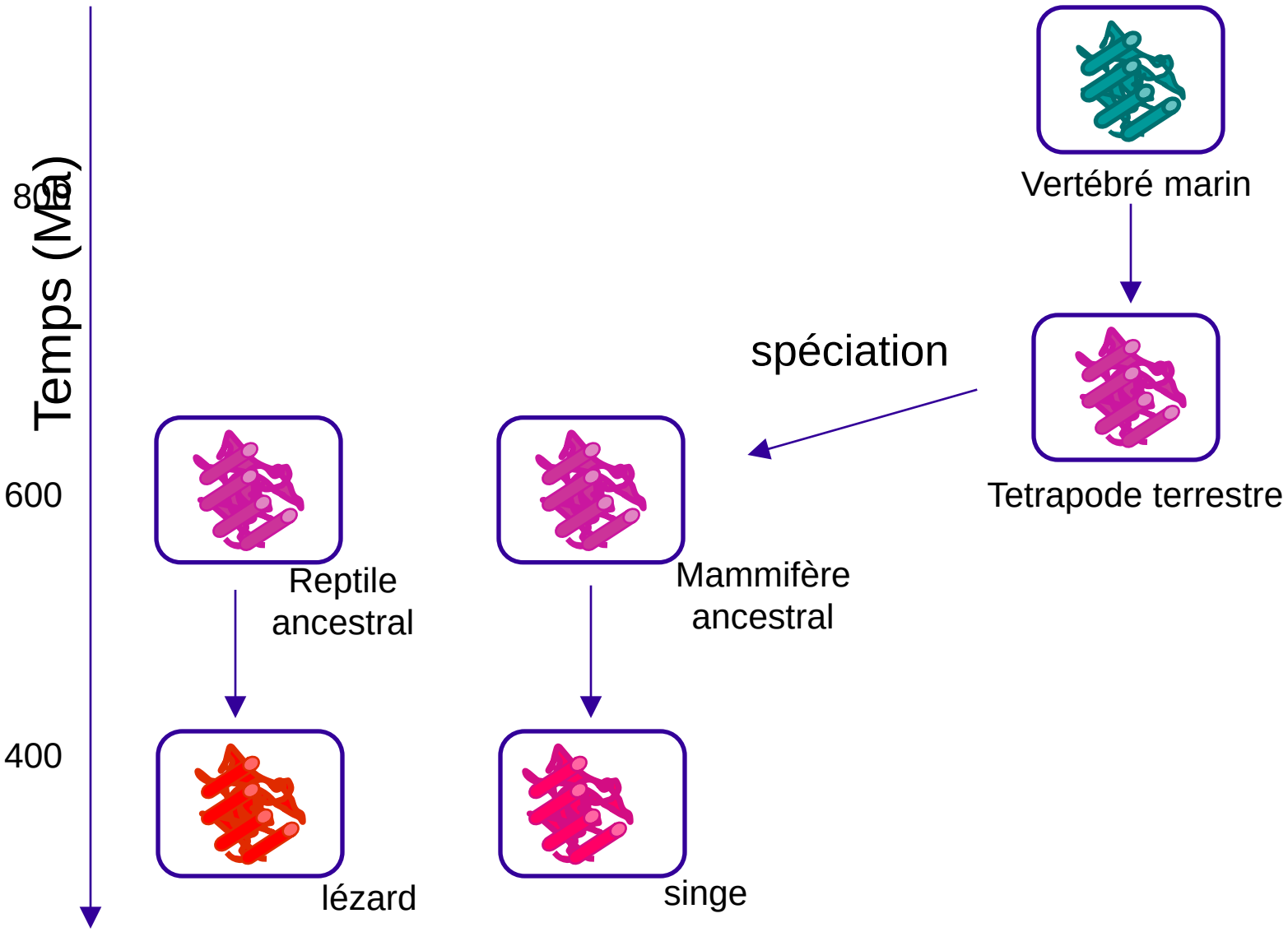
Fondements de l’annotation par homologie

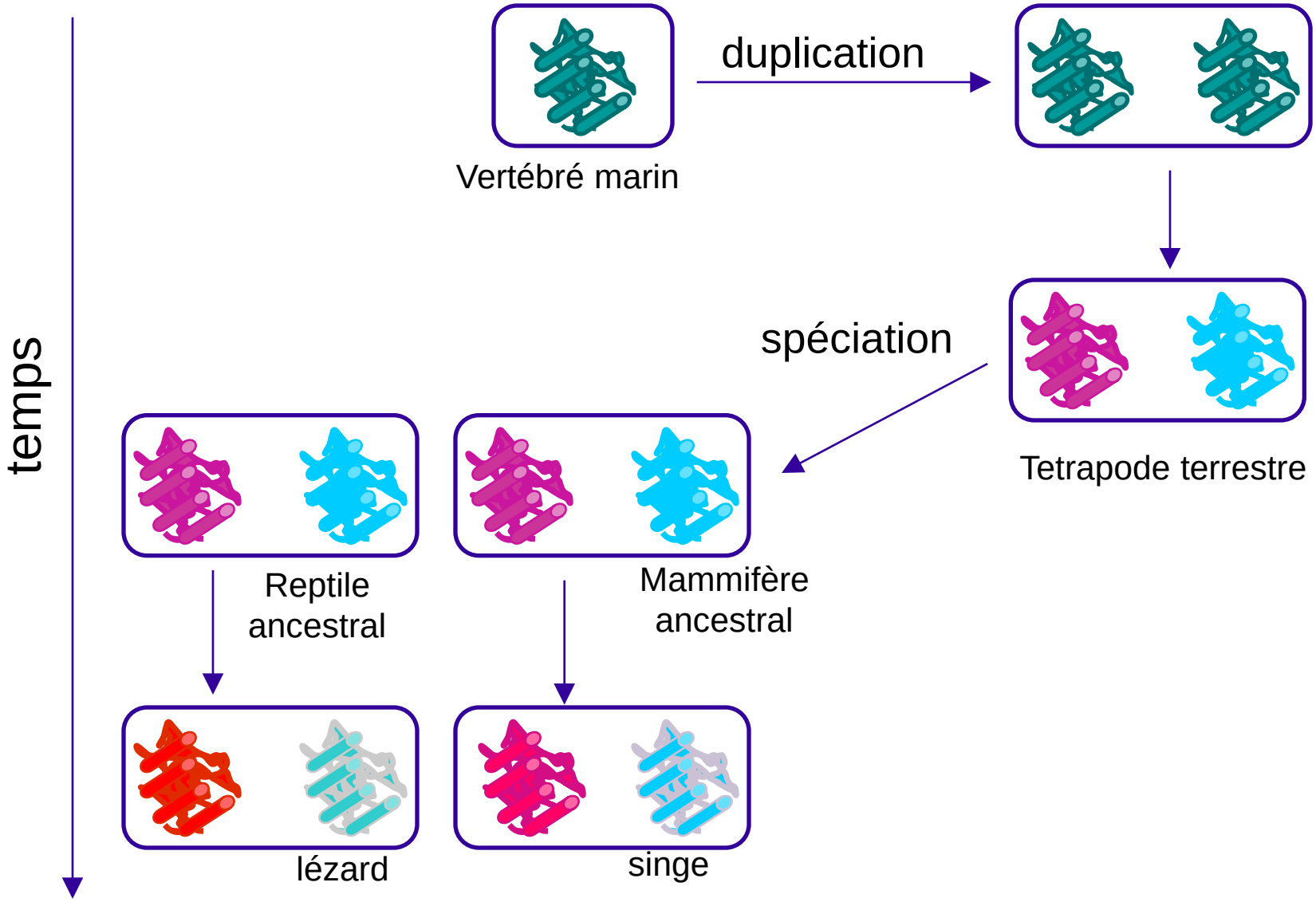
Comment les gènes apparaissent?

- Duplications
  - Duplication du génome entier ou polyploïdisation (plusieurs cas chez les eucaryotes, par exemple chez les téleostéens, principale classe de poissons)
  - Duplication d'un gène ou d'un groupe de gènes (fréquent)
  - Duplication d'un chromosome ou d'une partie (rare car délétère)
  - La duplication est suivie le plus fréquemment de la perte de gènes: 90% des gènes dupliqués à l'origine des vertébrés auraient été perdus depuis.
- Transferts horizontaux
  - Très important entre génomes procaryotes
  - Survient parfois de procaryote à eucaryote

Evolution d’un gène

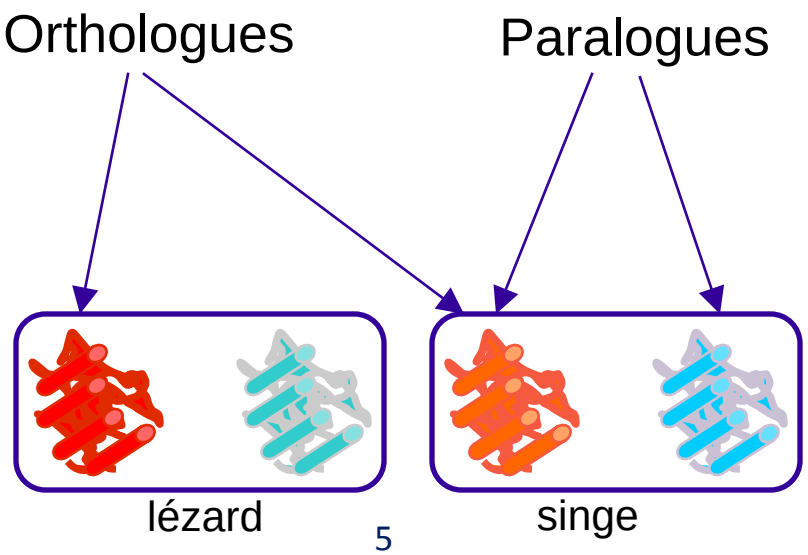


Apparition de nouveaux gènes par duplication



Paralogues et orthologues (Fitsh, 1970)

- ★ Homologues: gènes provenant d'un ancêtre commun
- ★ Paralogues: gènes homologues issus d'un phénomène de duplication
- ★ Orthologues: gènes homologues issus de la spéciation
- ★ Transfert horizontal: par endosymbiontes, etc. Fitsch a aussi introduit "xénologue" pour évoquer ce cas.

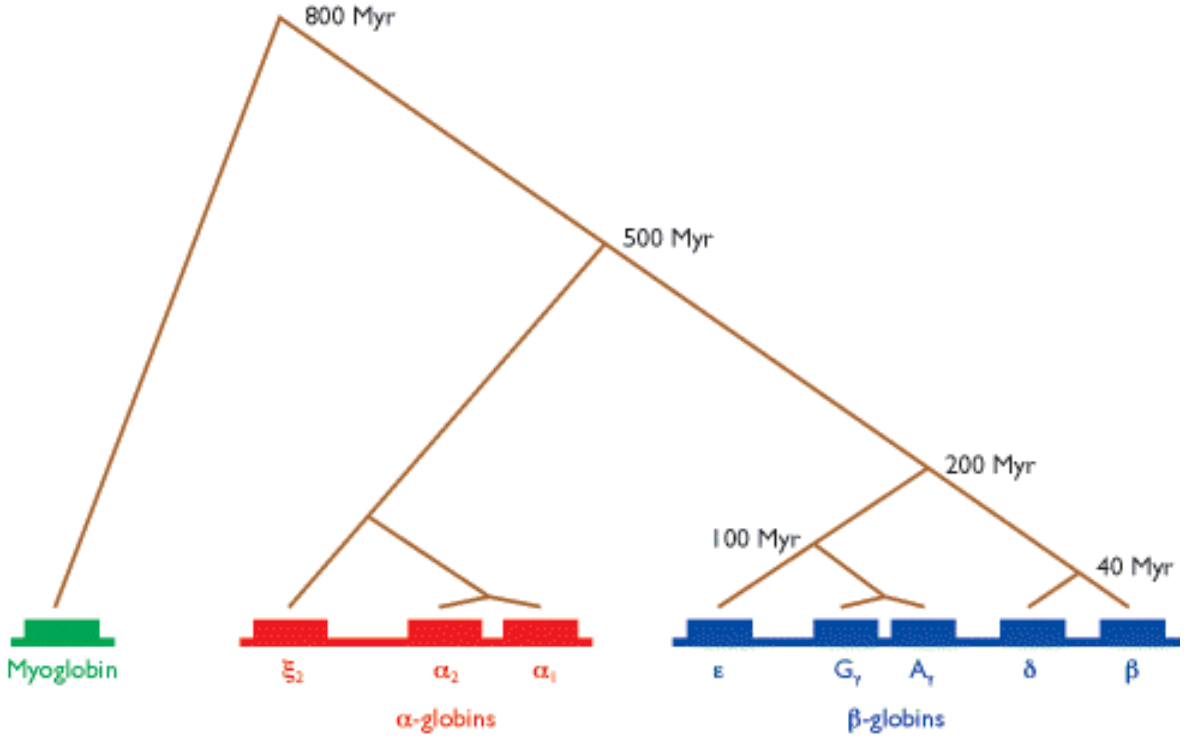


Fonction et homologie

- Homologie n'implique pas même fonction: par exemple l'aile de l'oiseau et le bras humain n'ont pas la même fonction
- Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.
- Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle *phénotypique*, mais peuvent exercer le même rôle dans une *voie* donnée.
- Les paralogues acquièrent rapidement des fonctions différentes

Exemple: les gènes de globine humains

★ Tous paralogues



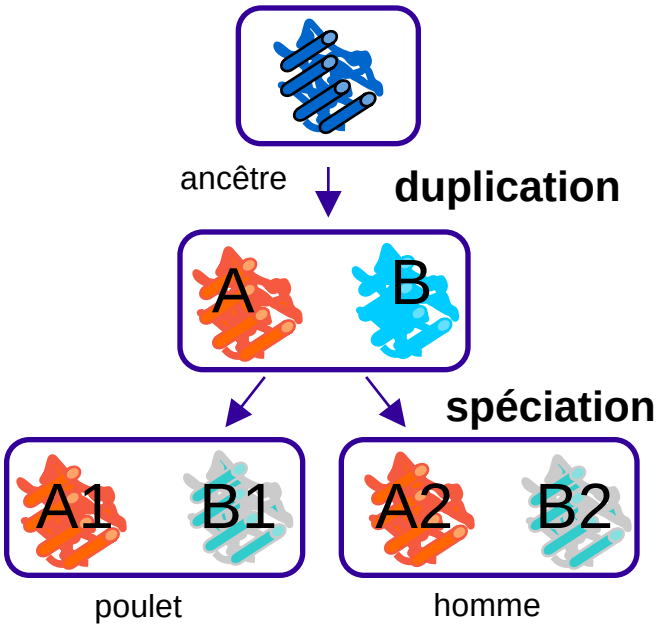
Les gènes se trouvent maintenant sur differents chromosomes: le gène de myoglobine est sur le chromosome 22, les gènes de α-globines sont sur le chromosome 16 et les gènes de β-globine sont sur le chromosome 11.

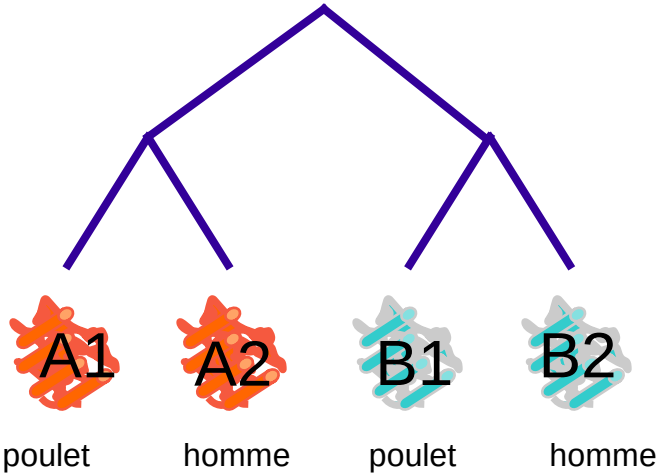
In « Genomes 2 », ©BIOS Scientific Publishers Ltd, 1999

Arbres avec paralogues et orthologues

★Admettons le schéma évolutif suivant (à gauche) ayant produit deux gènes paralogues présents chez tous les vertébrés.

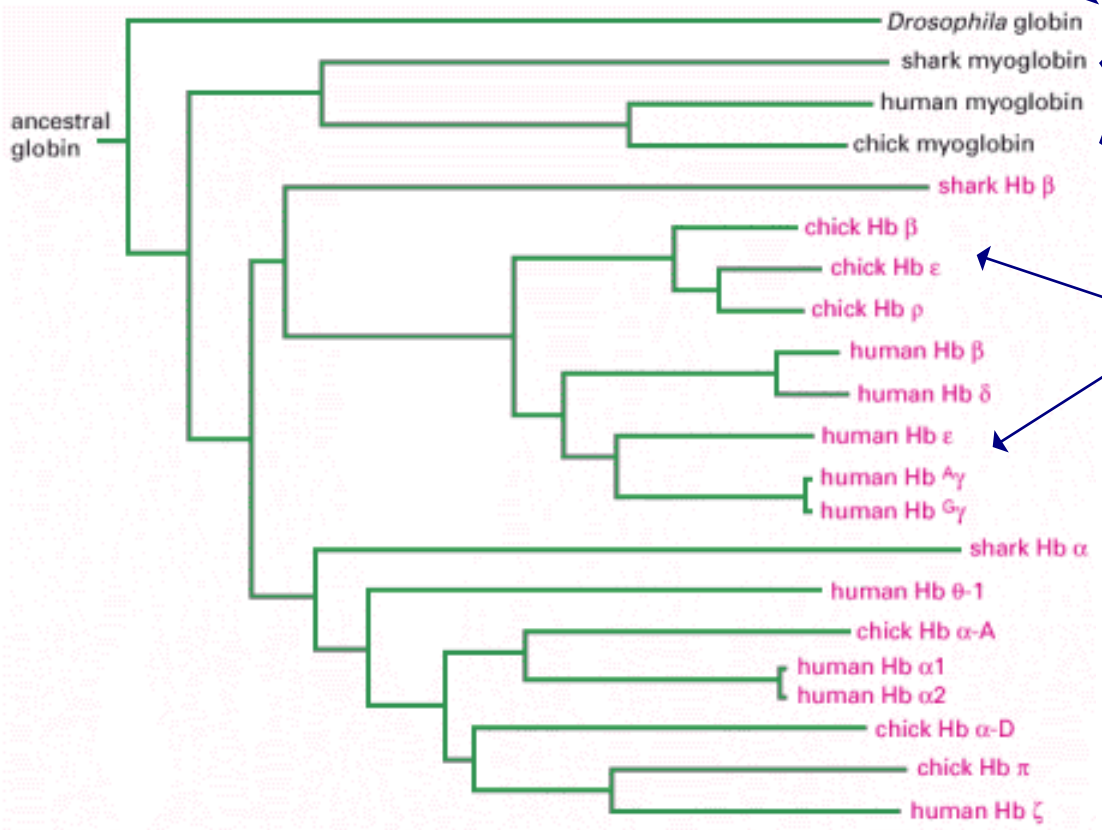
★ Etant donné que la duplication (ayant produit les paralogues) a eu lieu AVANT la spéciation (ayant produit les orthologues), les orthologues devraient être plus proches entre eux que les paralogues. L'arbre devrait donc ressembler à ceci:





Les gènes de globine chez # espèces

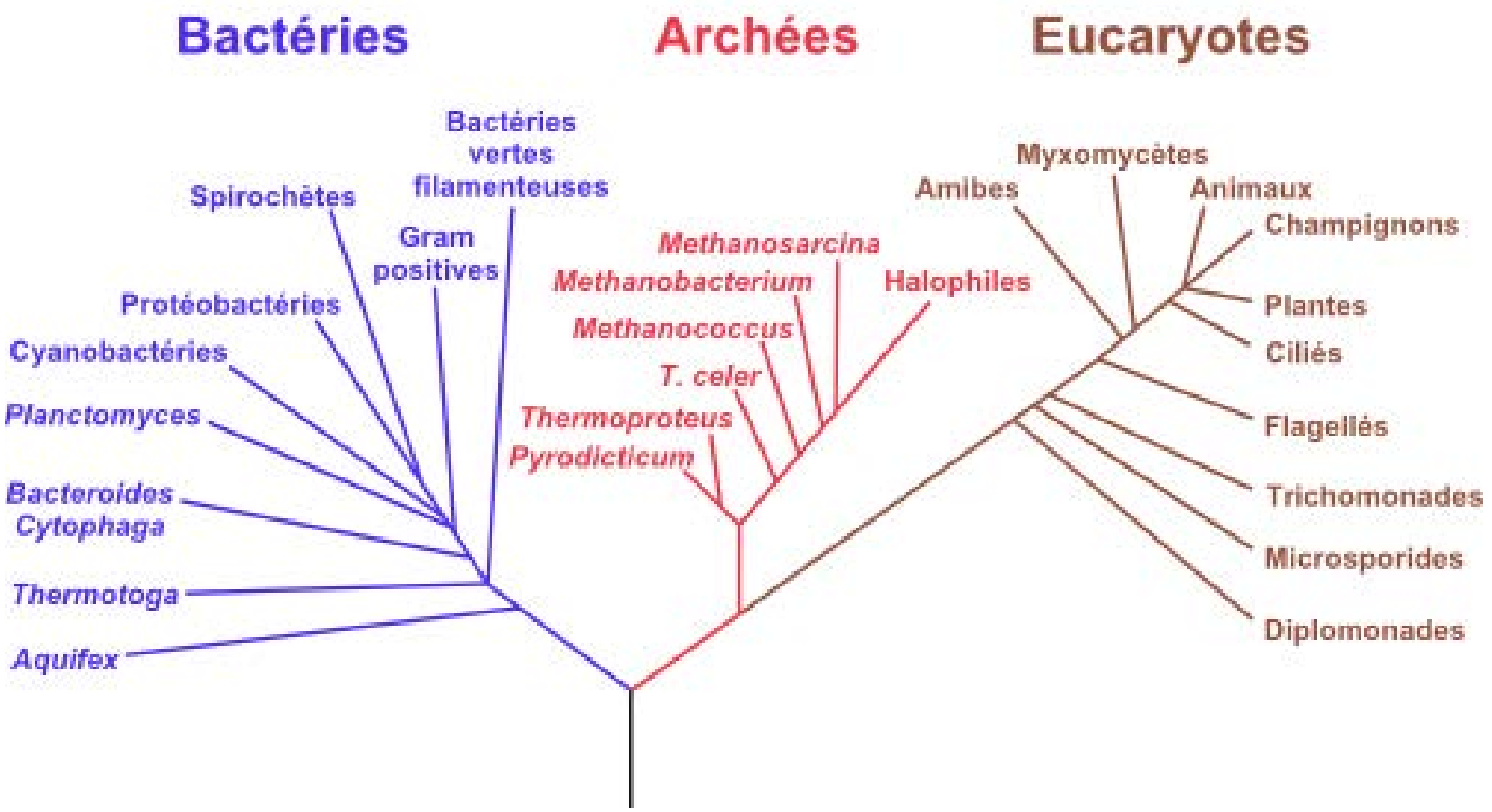
« Outgroup » = groupe extérieur. Indispensable pour placer la racine.

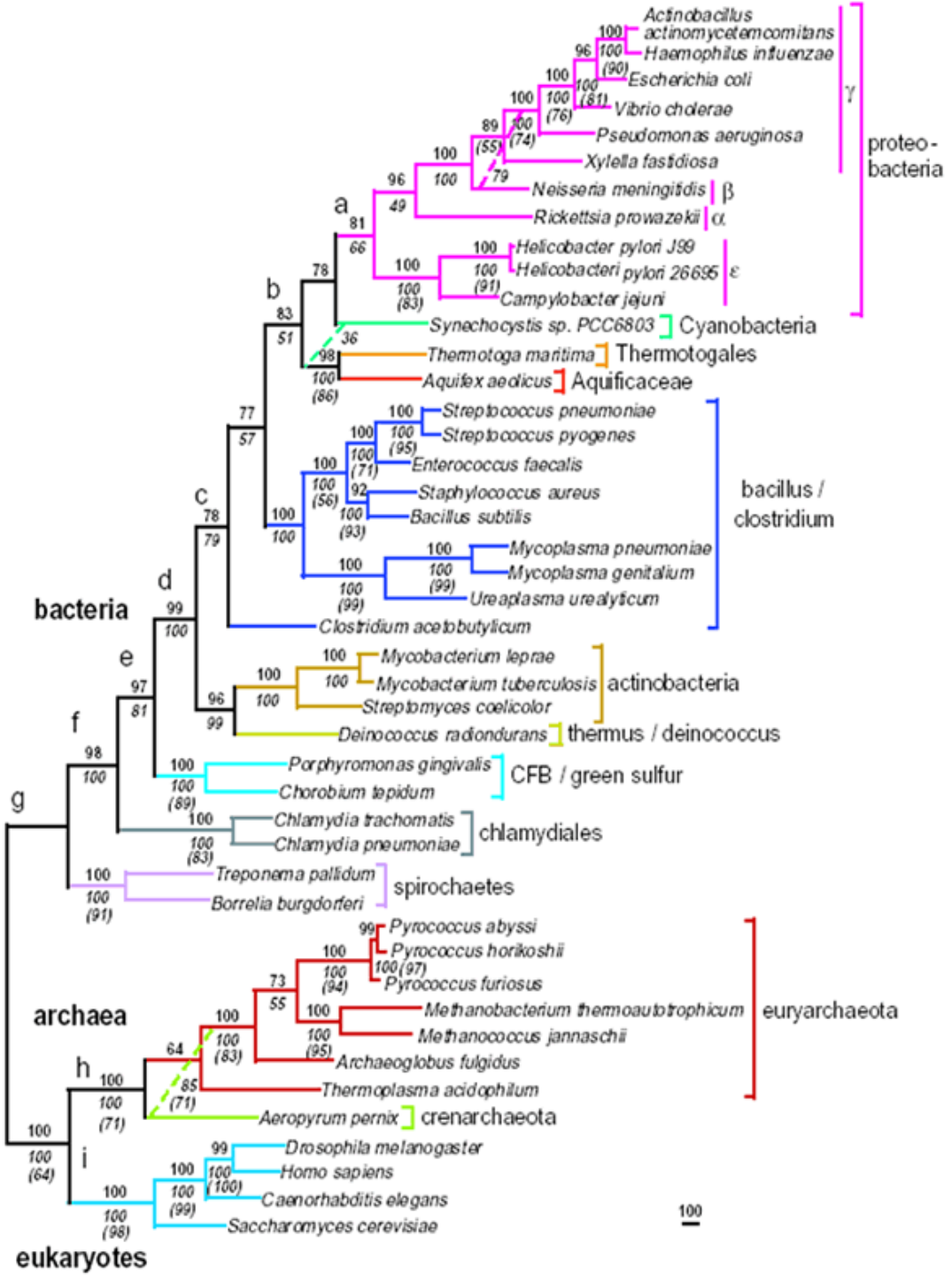


Myoglobines toutes orthologues.

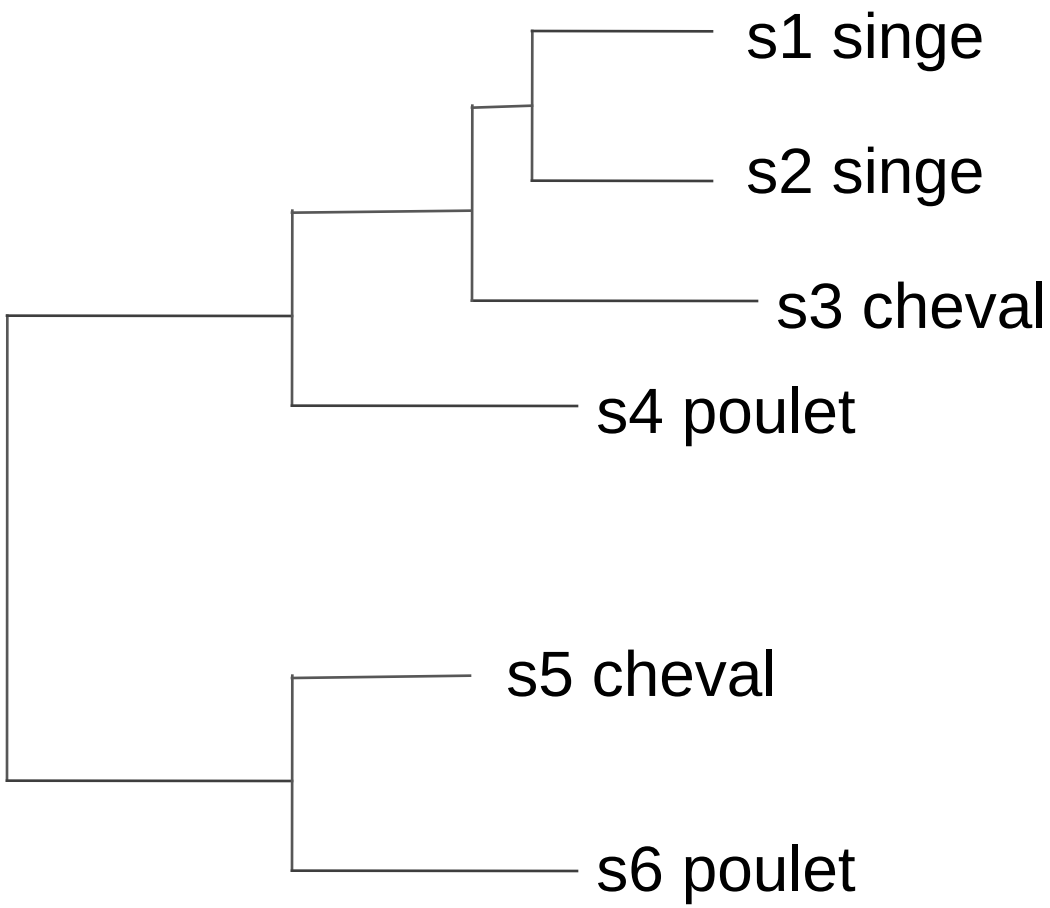
Globines epsilon homme et poulet non orthologues.
=> Deux évènements de duplication

Importance de la phylogénie

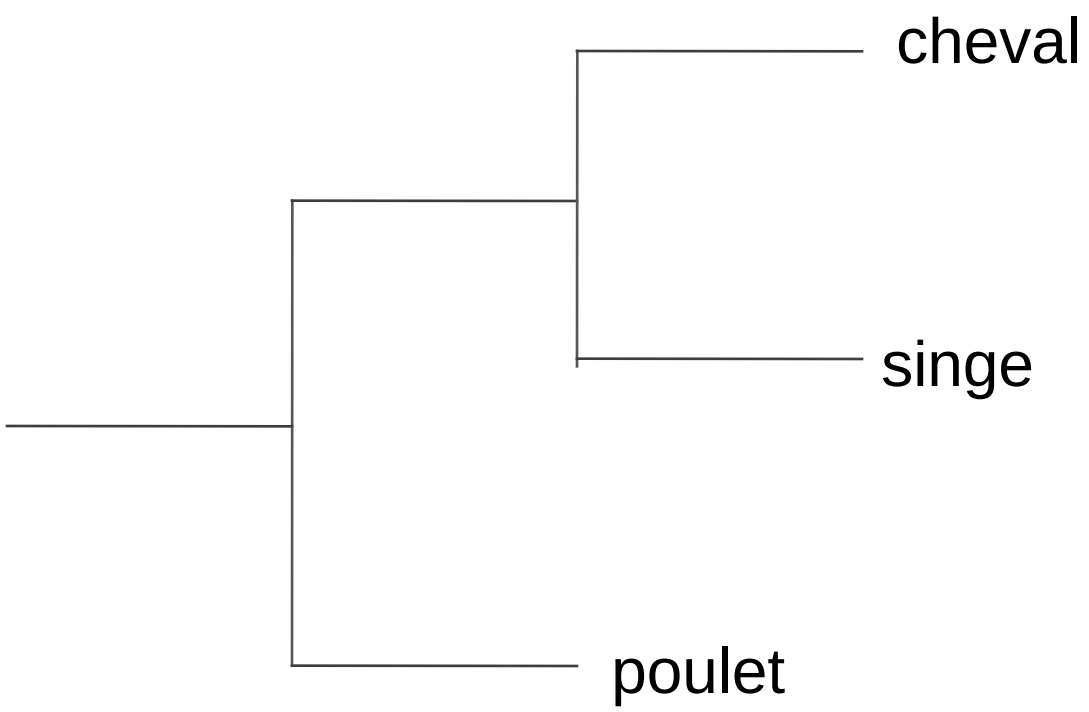



Actinobacillus actinomycetemcomitans
Haemophilus influenzae
Escherichia coli
Vibrio cholerae
Pseudomonas aeruginosa
Xylella fastidiosa
Neisseria meningitidis
Rickettsia prowazekii
Helicobacter pylori J99
Helicobacteri pylori 26695
Campylobacter jejuni
γ
β
α
ε
proteo-bacteria
Synechocystis sp. PCC6803
Cyanobacteria
Thermotoga maritima
Thermotogales
Aquifex aedicus
Aquificaceae
Streptococcus pneumoniae
Streptococcus pyogenes
Enterococcus faecalis
Staphylococcus aureus
Bacillus subtilis
Mycoplasma pneumoniae
Mycoplasma genitalium
Ureaplasma urealyticum
Clostridium acetobutylicum
bacillus / clostridium
Mycobacterium leprae
Mycobacterium tuberculosis
Streptomyces coelicolor
actinobacteria
Deinococcus radiondurans
thermus / deinococcus
Porphyromonas gingivalis
Chorobium tepidum
CFB / green sulfur
Chlamydia trachomatis
Chlamydia pneumoniae
chlamydiales
Treponema pallidum
Borrelia burgdorferi
spirochaetes
Pyrococcus abyssi
Pyrococcus horikoshii
Pyrococcus furiosus
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Archaeoglobus fulgidus
Thermoplasma acidophilum
euryarchaeota
Aeropyrum pernix
crenarchaeota
Drosophila melanogaster
Homo sapiens
Caenorhabditis elegans
Saccharomyces cerevisiae
bacteria
archaea
eukaryotes
a
b
c
d
e
f
g
h
i
100

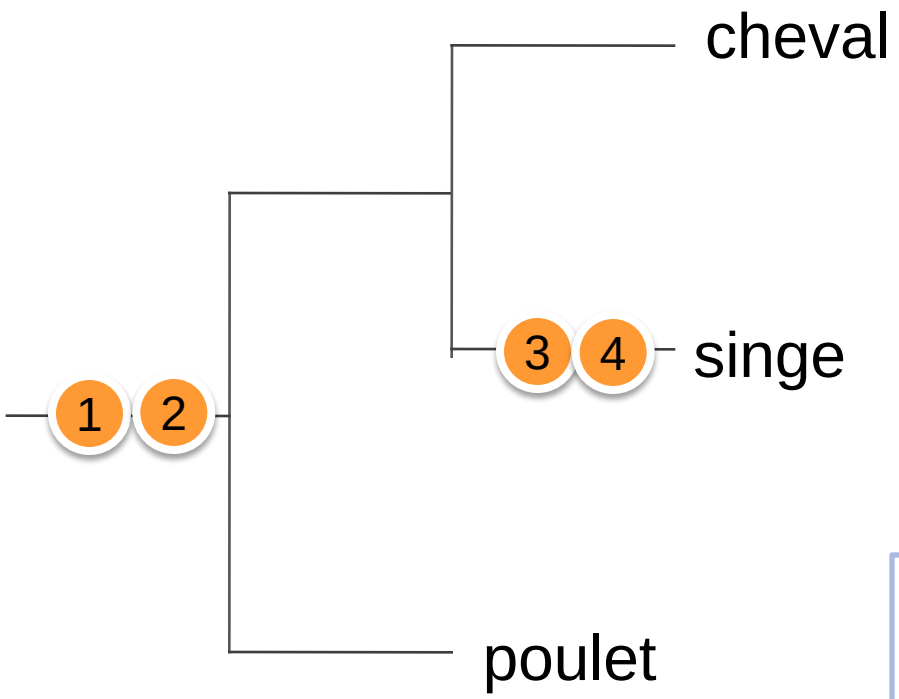
Exercice



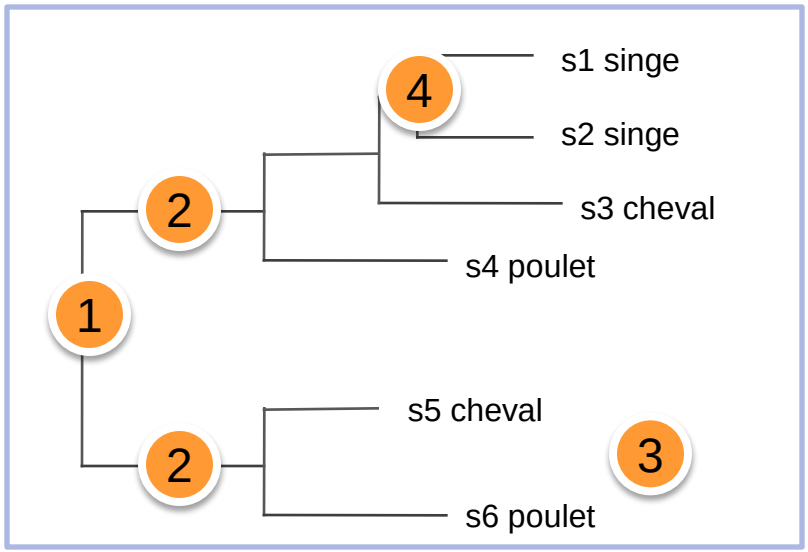
D’abord faire l’arbre des espèces



Puis placer les évènements



1. Apparition du gène ancestral P+C+S
2. Duplication chez ancêtre mamm+oiseaux
3. Perte d'une copie chez le singe
4. Duplication « récente » chez le singe



Vaut-il mieux comparer les protéines ou l'ADN pour rechercher des homologues d'une séquence?

- ★ La meilleure façon de détecter des similitudes entre séquences est généralement la *comparaison au niveau protéique.*
  1. Il existe 20 aa contre 4 bases. La probabilité de trouver une "lettre" donnée par hasard est donc plus importante pour les bases.
  2. Plusieurs codons produisent le même aa. 134 / 549 substitutions de bases sont synonymes. Les séquences protéiques sont plus informatives.
  3. La raison principale est en fait l'existence d'outils de comparaison plus puissants pour les aa: utilisation des propriétés physicochimiques ou des substitutions observées dans l'évolution. Même lorsque les aa sont différents, on est capable de retrouver des similitudes. On en est tout à fait incapable au niveau des bases.
- ★ Il existe en fait des cas où la séquence d'ADN est plus conservée que la séquence protéique, ce qui enlève du poids à l'argument 1
- ★ Les comparaisons avec les séquences protéiques ne permettent de détecter que les régions codantes. Evidemment, on utilisera toujours la séquence ADN/ARN pour analyser ce qui n'est pas traduit!

L’analyse des domaines et de l’alignement multiple

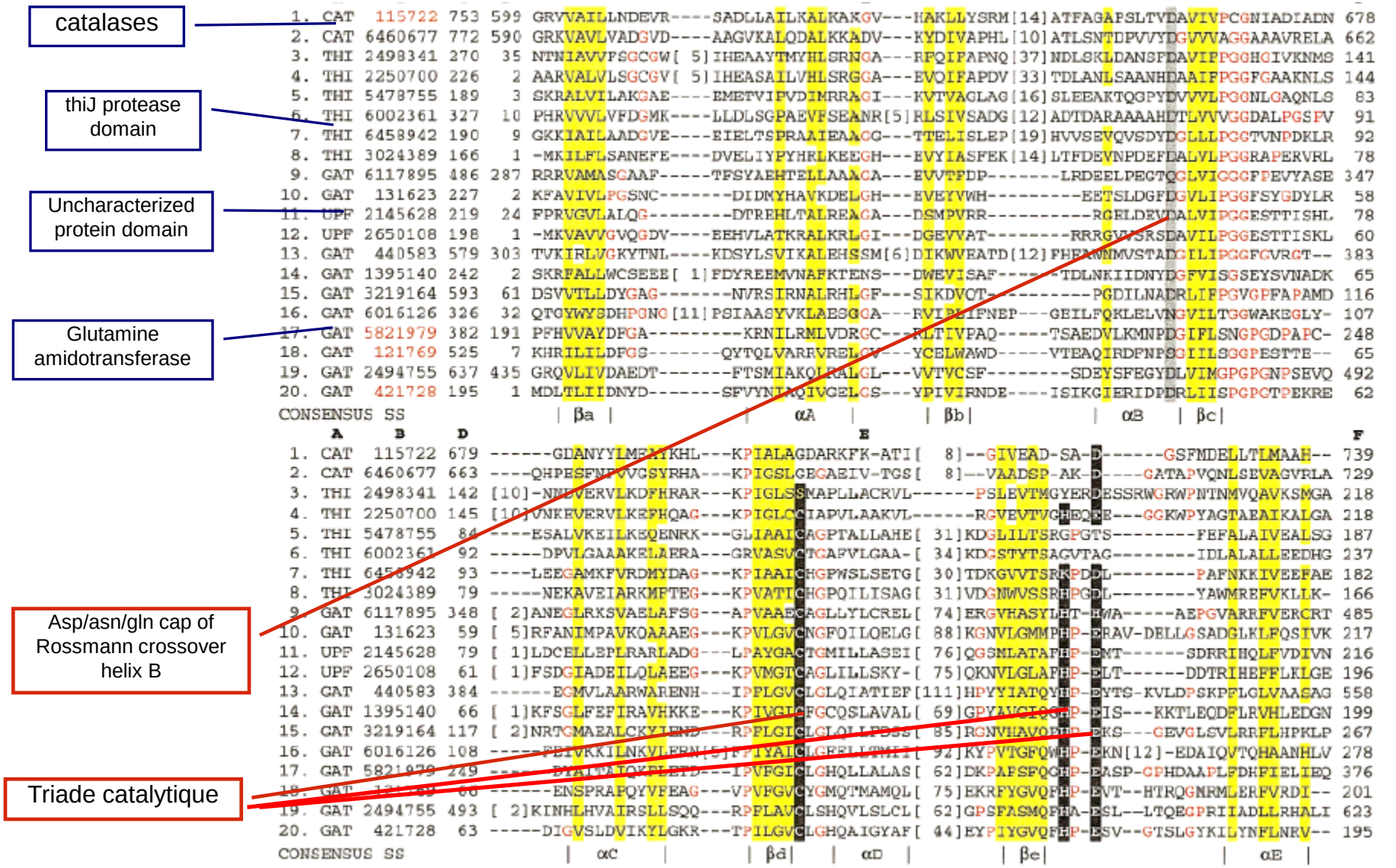
Alignement multiple

- **Pourquoi réaliser un alignement multiple?**
- L'alignement multiple révèle des fonctions que l'on ne pouvait pas visualiser en comparant 2 séquences

  - Identifier les positions et les acides aminés importants.
  - Visualiser les domaines
  - Distinguer paralogues et orthologues
  - Etablir la phylogénie des séquences, et même parfois des organismes
  - Comme une aide à la modelisation: Les algorithmes de prediction de structures secondaires exploitent beaucoup mieux les alignements multiples. Connaitre les aminoacides permis à telle ou telle position facilite l'inférence 3D.

Exemple d'analyse d'alignement multiple: les Glutamine Aminotransferases (GAT)

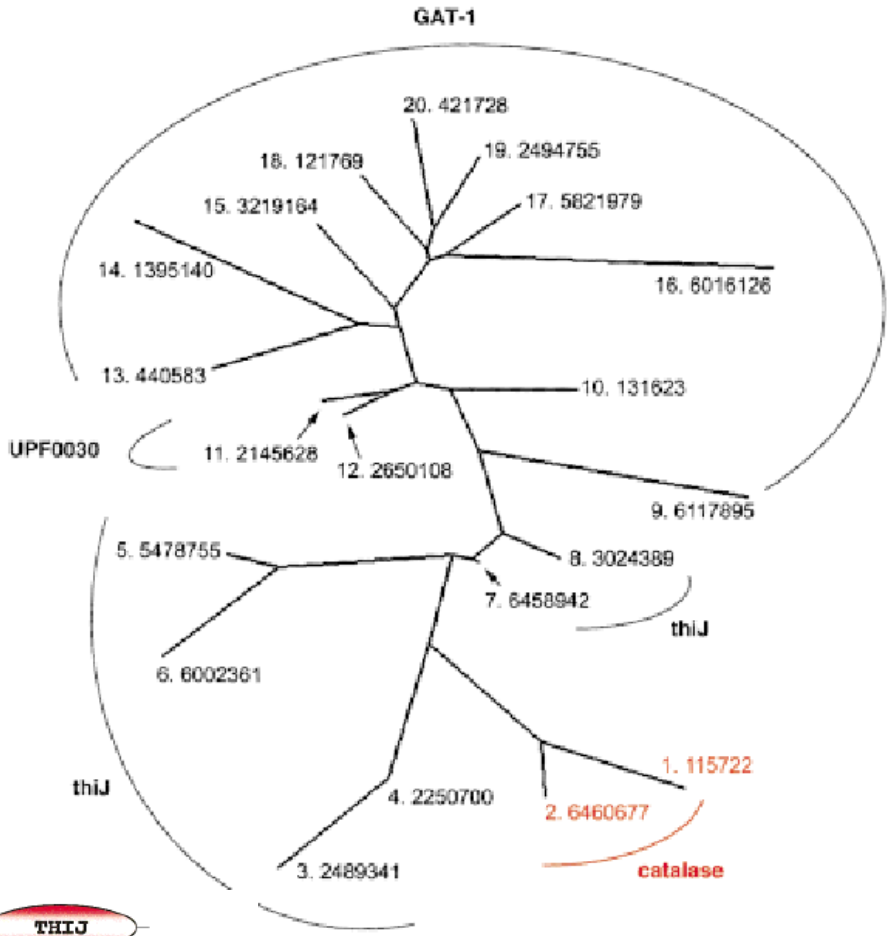
Horvath & Grishin, Proteins, 2001

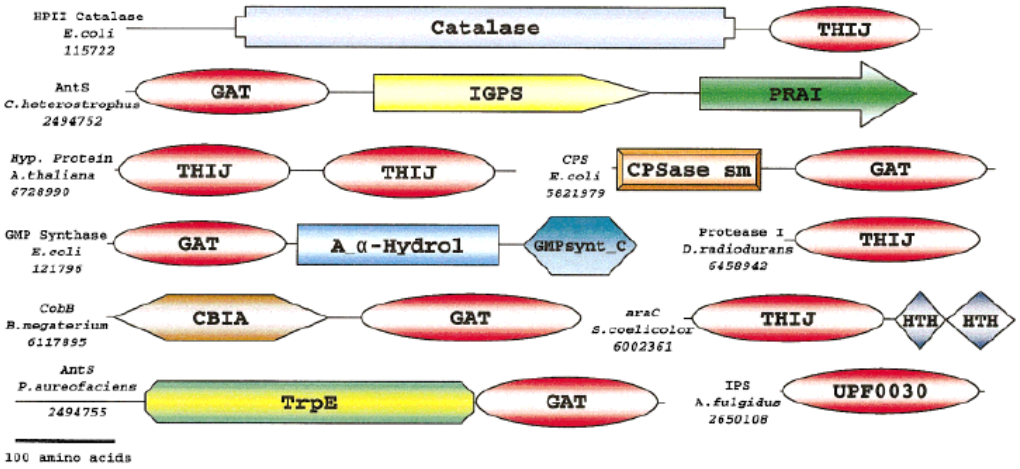


1, Escherichia coli HPII catalase; 2, D. radiodurans HPII catalase; 3, Danio rerioES1; 4, Homo sapiensKNP-I a; 5, Rattus norvegicusSP22; 6, S. coelicolorAraC; 7, D. radioduransprotease I; 8, P. furiosusPfpI protease I; 9, S. coelicolor CobB; 10, Bacilis subtilisFGAM synthase I; 11, Mycobacterium leprae amidotransferase hisH; 12, A. fulgidus imidazole glycerol-phosphate synthase subunit H; 13, Saccharomyces cerevisiaeCTP synthetase; 14, Acinetobacter sp. aniline dioxygenase; 15, Arabidopsis thaliana glutamine amidotransferase; 16, A. thaliana g-glutamyl hydrolase precursor; 17, E. coli carbamoyl phosphate synthetase; 18, E. coliGMP synthase, P. aureofaciens AntS, S. sulfataricus AntS.

Mobilité des domaines entre protéines différentes

Exemple des Glutamine Aminotransferases (GAT):





… de l'importance de séparer les domaines pour l'analyse

Banques de domaines

✦ Prosite

Expression régulière

SIGMA70_2, PS00716; **Sigma-70 factors family signature 2** (PATTERN)

Consensus pattern: **[STN]-x(2)-[DENQ]-[LIVMT]-[GAS]-x(4)-[LIVMF]-[PSTG]-x(3)-[LIVMA]-x-[NQR]-[LIVMA]-[EQH]-x(3)-[LIVMFWK]-x(2)**

✦ PFAM

There are 346 sequences with the following architecture: DUF1785, PAZ, Piwi

AGOL_ARATH [Arabidopsis thaliana (Mouse-ear cress)] Argonaute-like protein At2g27880 (997 residues)

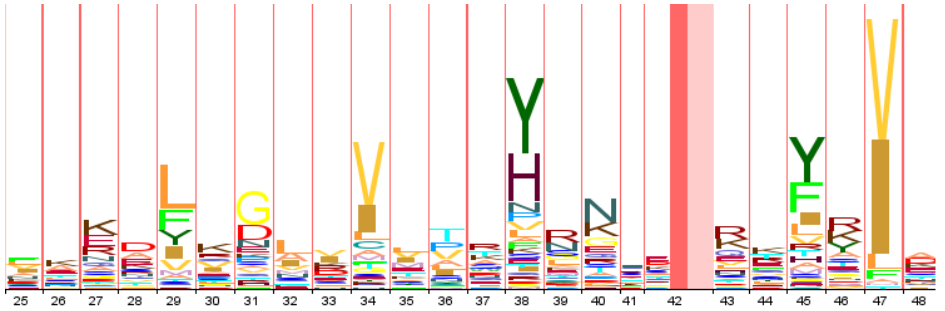
Show all sequences with this architecture.

There are 232 sequences with the following architecture: PAZ, Piwi

PIWL2_HUMAN [Homo sapiens (Human)] Piwi-like protein 2 (973 residues)

Show all sequences with this architecture.

HMM logo



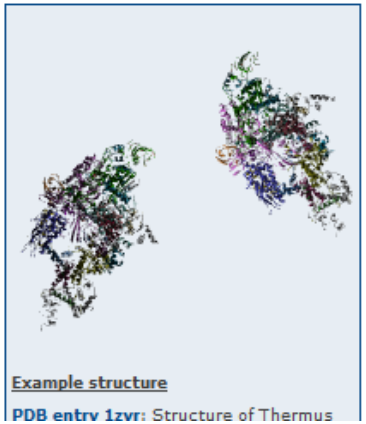
Informations précieuses présentes dans les banques de domaines

Sigma-70 region 2 Add annotation

Region 2 of sigma-70 is the most conserved region of the entire protein. All members of this class of sigma-factor contain region 2. The high conservation is due to region 2 containing both the -10 promoter recognition helix and the primary core RNA polymerase binding determinant. The core binding helix, interacts with the clamp domain of the largest polymerase subunit, beta prime [1,2]. The aromatic residues of the recognition helix, found at the C-terminus of this domain are though to mediate strand separation, thereby allowing transcription initiation [1,2].

Literature references

1. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA; , Mol Cell 2002;9:527-539.: Structure of the bacterial RNA polymerase promoter specificity sigma subunit. PUBMED:11931761
2. Malhotra A, Severinova E, Darst SA; , Cell 1996;87:127-136.: Crystal structure of a sigma 70 subunit fragment from E. coli RNA polymerase. PUBMED:8858155

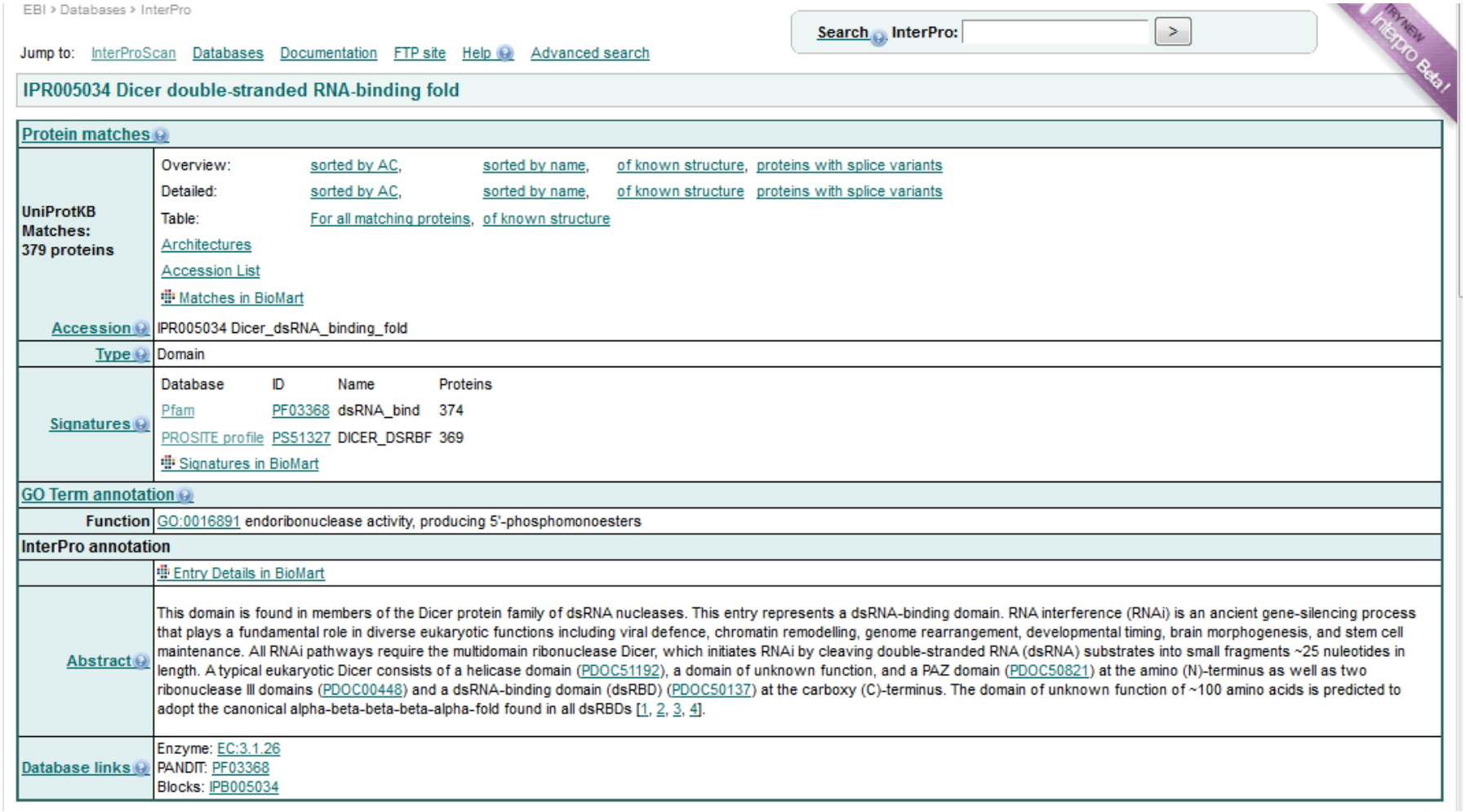


Example structure

PDB entry 1zyr: Structure of Thermus

Exemple d'annotation de domaine dans PFAM

Interpro: la banque de domaines intégrée de l'EBI



EBI > Databases > InterPro

Search InterPro: >

Try new InterPro Beta!

Jump to: InterProScan Databases Documentation FTP site Help Advanced search

IPR005034 Dicer double-stranded RNA-binding fold

Protein matches

UniProtKB Matches: 379 proteins

Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants

Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants

Table: For all matching proteins, of known structure

Architectures

Accession List

Matches in BioMart

Accession IPR005034 Dicer_dsRNA_binding_fold

Type Domain

Signatures

Database	ID	Name	Proteins
Pfam	PF03368	dsRNA_bind	374
PROSITE profile	PS51327	DICER_DSRBF	369

Signatures in BioMart

GO Term annotation

Function GO:0016891 endoribonuclease activity, producing 5'-phosphomonoesters

InterPro annotation

Entry Details in BioMart

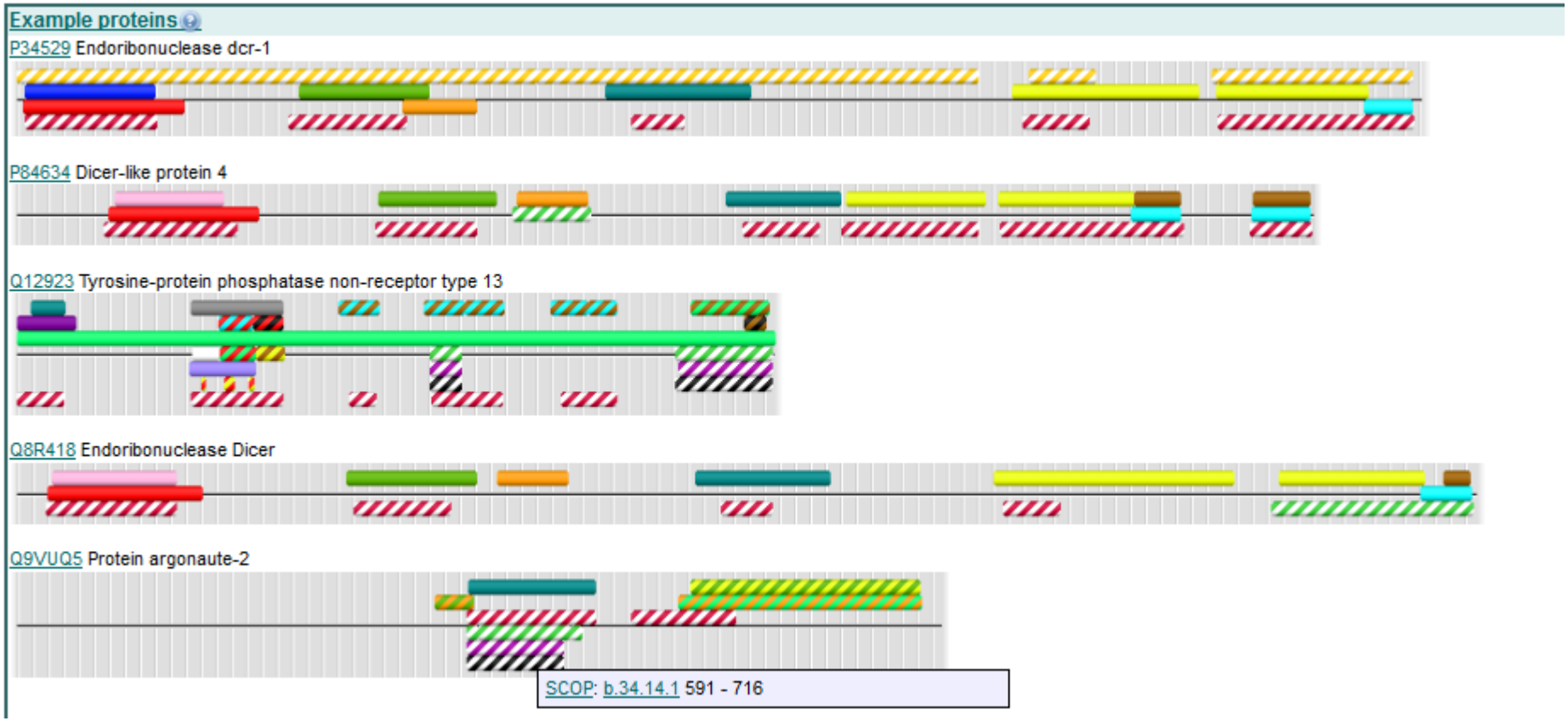
Abstract

This domain is found in members of the Dicer protein family of dsRNA nucleases. This entry represents a dsRNA-binding domain. RNA interference (RNAi) is an ancient gene-silencing process that plays a fundamental role in diverse eukaryotic functions including viral defence, chromatin remodelling, genome rearrangement, developmental timing, brain morphogenesis, and stem cell maintenance. All RNAi pathways require the multidomain ribonuclease Dicer, which initiates RNAi by cleaving double-stranded RNA (dsRNA) substrates into small fragments ~25 nuleotides in length. A typical eukaryotic Dicer consists of a helicase domain (PDOC51192), a domain of unknown function, and a PAZ domain (PDOC50821) at the amino (N)-terminus as well as two ribonuclease III domains (PDOC00448) and a dsRNA-binding domain (dsRBD) (PDOC50137) at the carboxy (C)-terminus. The domain of unknown function of ~100 amino acids is predicted to adopt the canonical alpha-beta-beta-beta-alpha-fold found in all dsRBDs [1, 2, 3, 4].

Database links

Enzyme: EC:3.1.26
PANDIT: PF03368
Blocks: IPB005034

Interpro: la banque de domaines intégrée de l'EBI



Chaque rectangle coloré: un domaine identifié par une des banques (par exemple: banque SCOP)