GBSA

réunion du 1 juin 2016

présentation de Claudine Landès Devauchelle

IRHS
Institut de Recherche en Horticulture et Semences

équipe bioinformatique

INRA - AgroCampus Ouest - Université Angers

INRA SCIENCE & IMPACT
AGRO CAMPUS OUEST
université angers

Introduction

Equipe bioinformatique IRHS (une dizaine de membres)

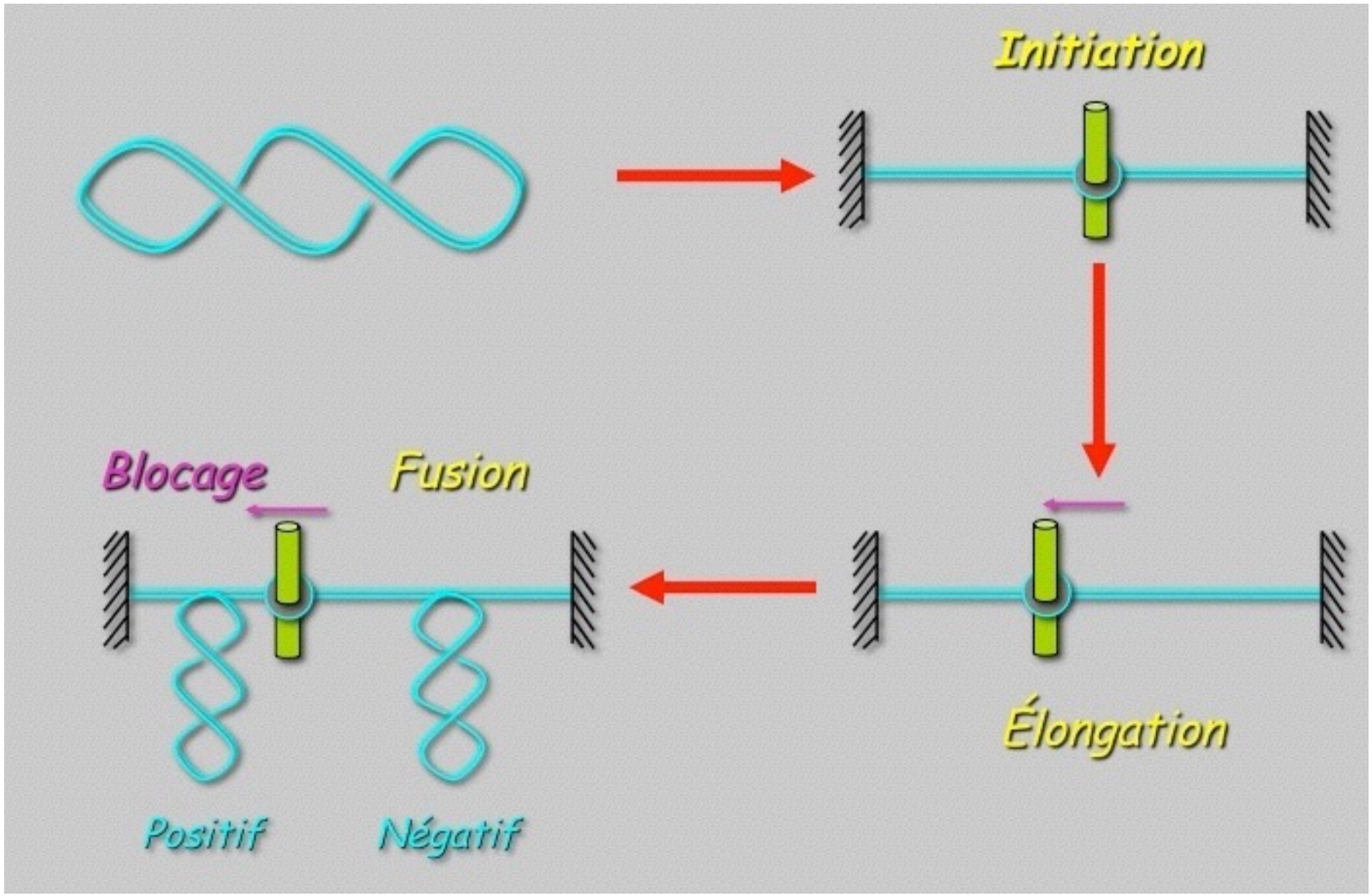
- recherche : Sébastien Aubourg/Claudine Landès
- service : ELVIS / ANANdb / CRB / outils d'analyses
- formation : soutien aux équipes de l'IRHS (projets)

Présentation

Classification de séquences protéiques divergentes (plusieurs milliers de séquences homologues) grâce à des méthodes à base de k-mers

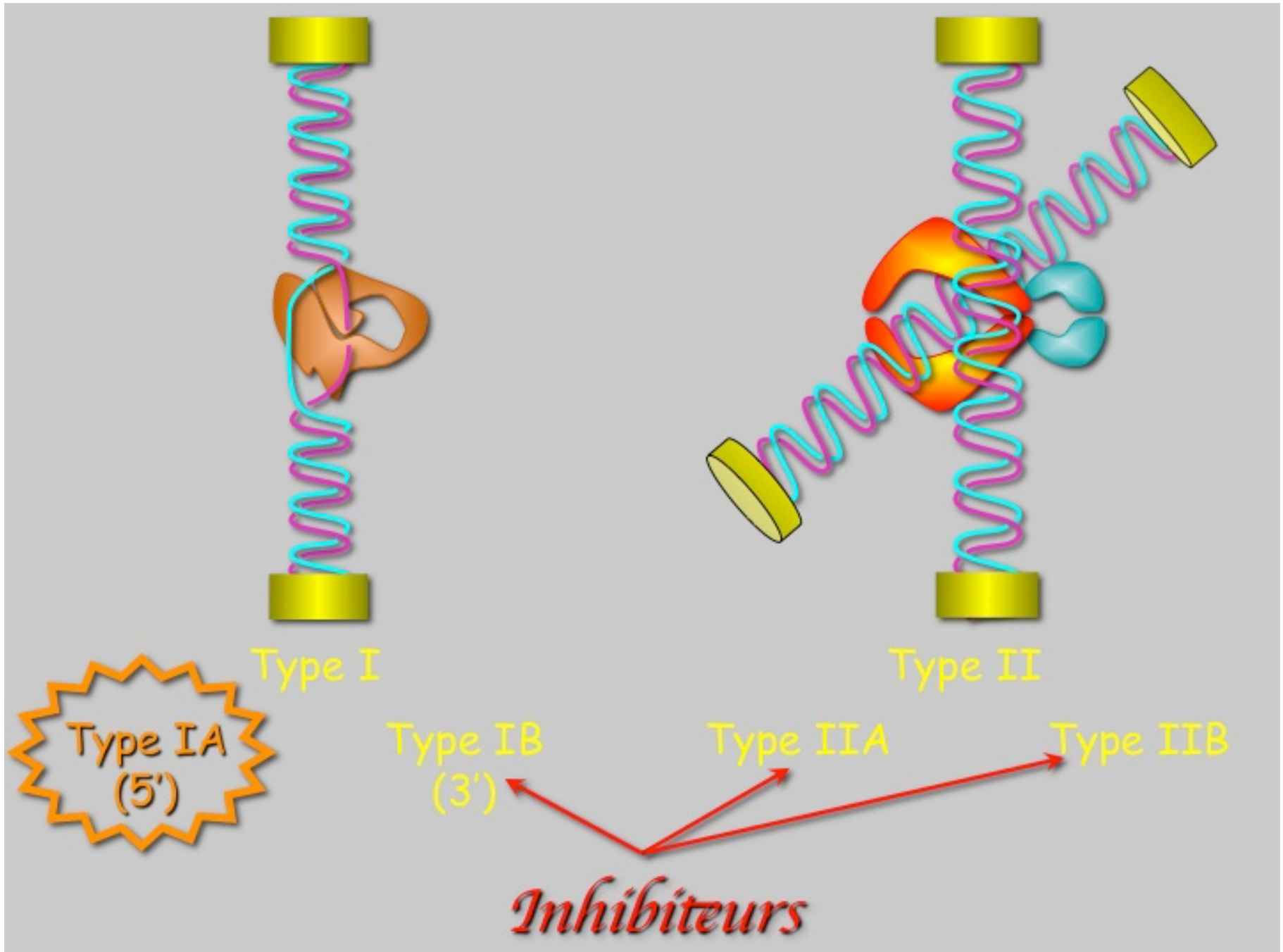
Classification sans alignement

Rôle des topoisomérases



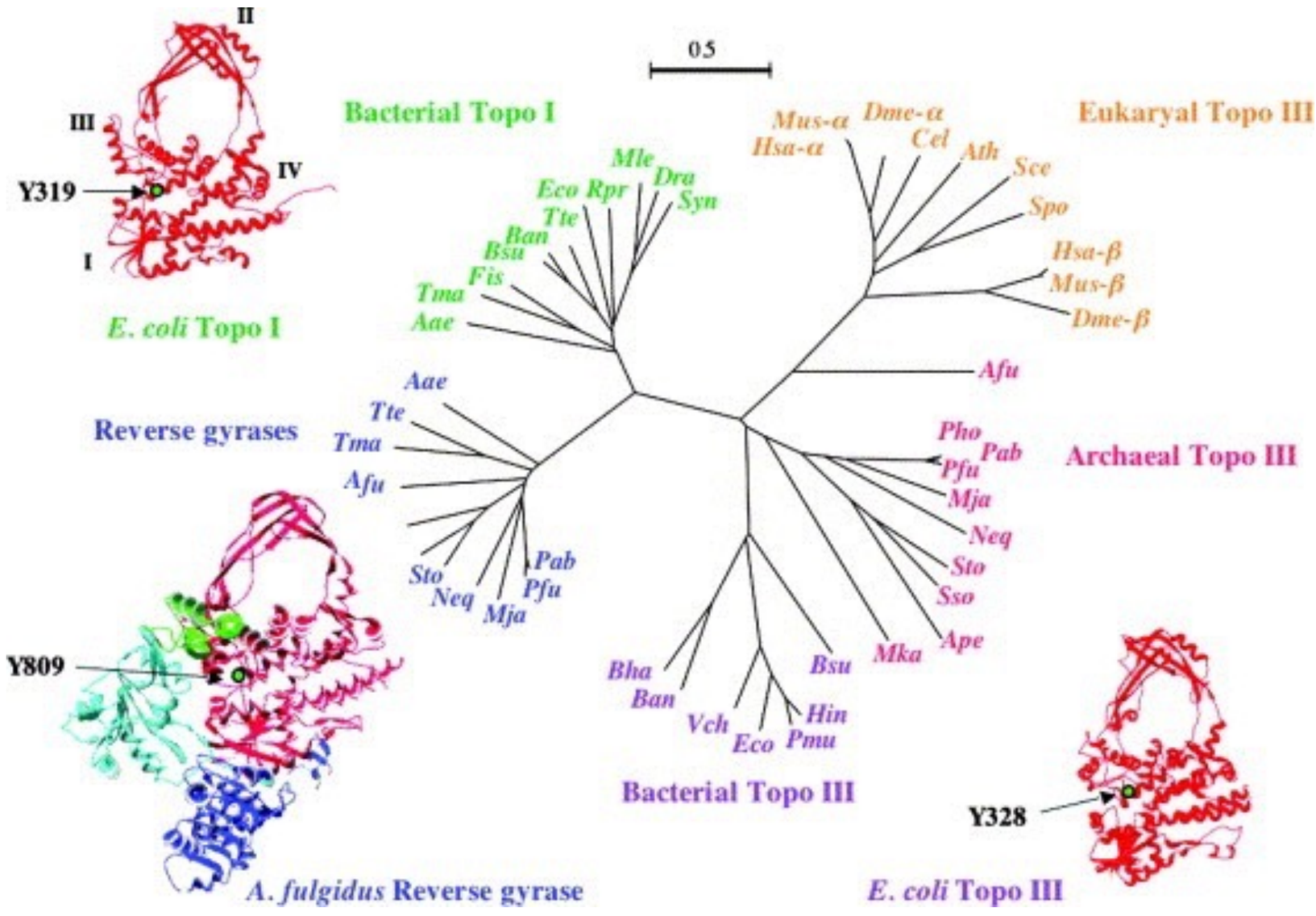
Marc Nadal, IGM, 2010

Classes de topoisomérases



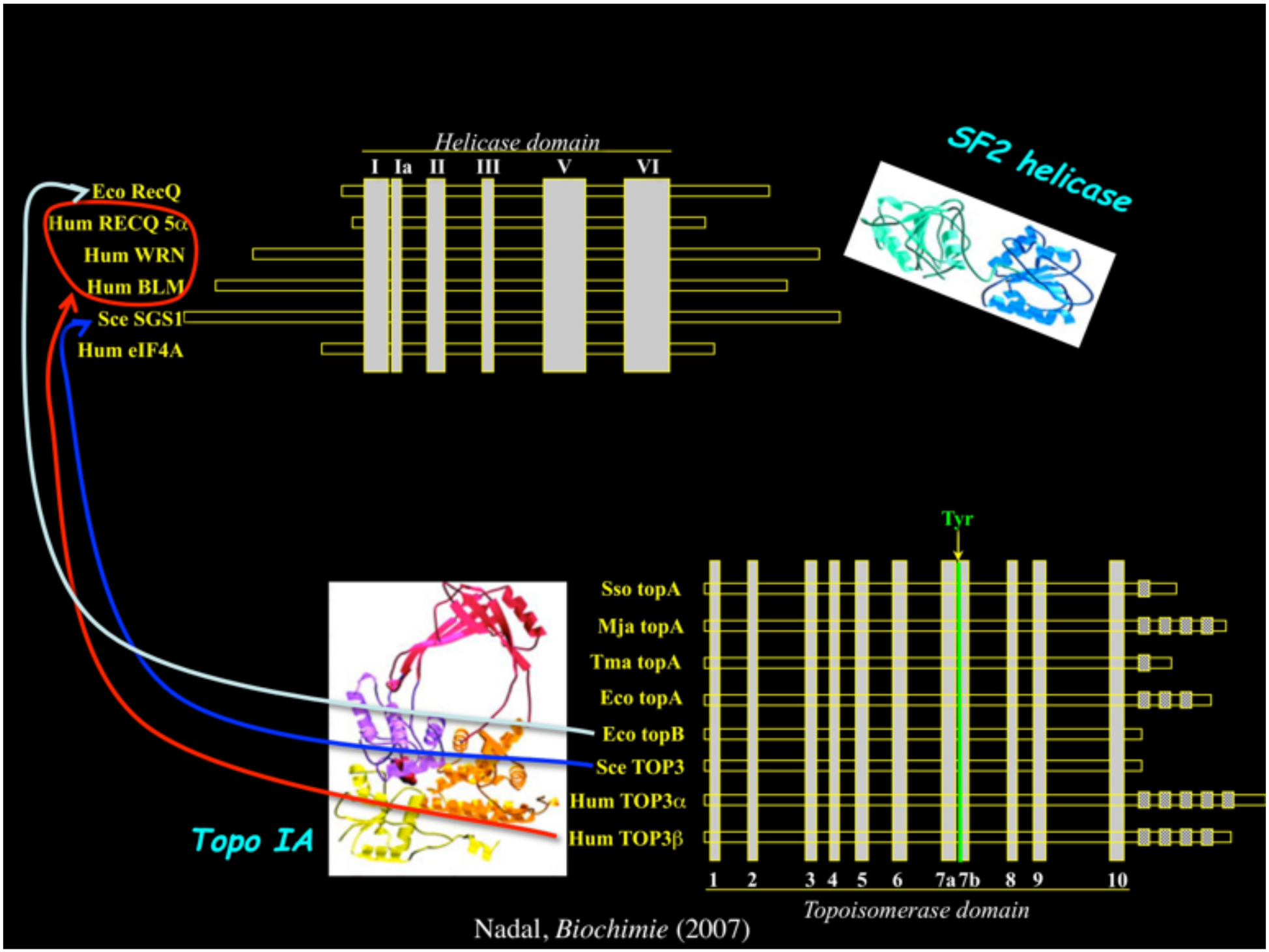
Marc Nadal, IGM, 2010

Sous-classes de type IA

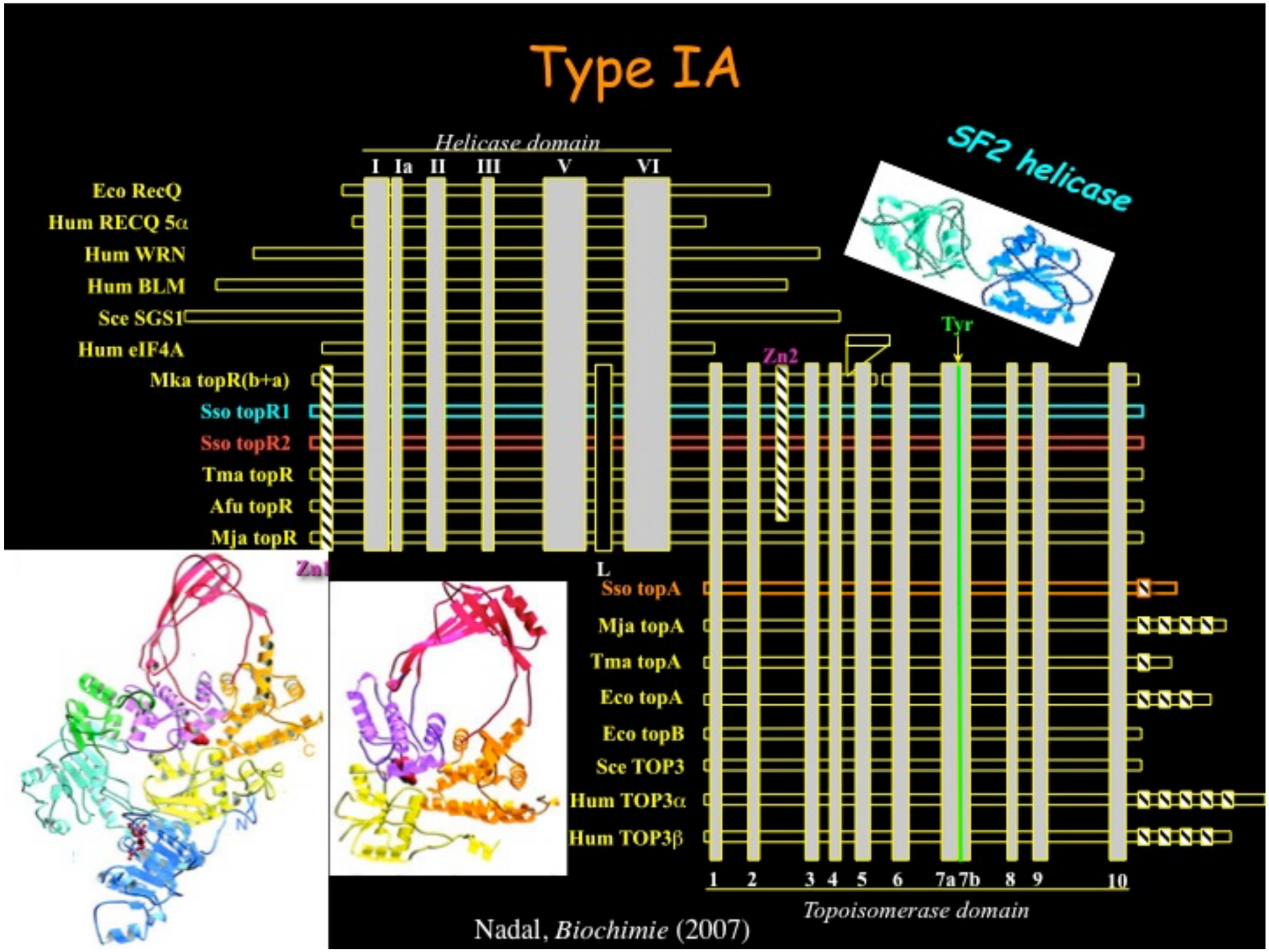


Michel Duguet, IGM, 2006

Topoisomérases IA



Les reverses gyrases



Séquences divergentes

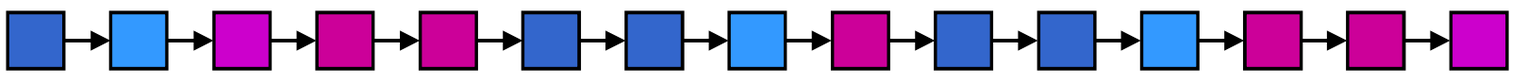
- taille variable : 800 à 1600 acides aminés
- dans tous les règnes du vivant
- divergentes entre les blocs conservés
- extrémité C-terminale variable (+/- doigts Zn)
- études de phylogénomiques possibles mais…

- taille variable : 800 à 2000 acides aminés
- séquences nombreuses et divergentes
- alignements multiples difficilement exploitables
- peu de caractères homologues pour l'analyse

Classification sans alignement

Point de départ

- Une séquence peut être vue comme une suite d'états
  - Au sens des chaînes de Markov cachées
  - Les séquences d'ADN possèdent 4 états
- Il est possible de diminuer le nombre d'états
  - Ex : ne distinguer que les purines (A, G) et les pyrimidines (T, C)



agtccaagcaagcct

123441124112443

...

décodage

αα β β β αααβαααβ β β

Maude Pupin, LIFL, 2007

Qu'est-ce que le décodage local à l'ordre N ?

- Il est également possible d'augmenter le nombre d'états
  - Les 4 nucléotides peuvent être vus comme une projection d'un alphabet plus grand
- Le décodage local à l'ordre N détermine les nouveaux états associés à chaque position d'une séquence en fonction des N nucléotides entourant chaque position de façon quelconque
  - Des classes d'équivalence sont définies entre les positions
- Un cas particulier : le décodage local maximal
  - Recherche du nombre maximal d'états

1. Exemple de segments regroupés par le décodage sur 23 RG (N=7)

```
SsoR2_T_A_Ss     99     SFIMSA P TGLGKT
ApeR1_T_A_Ap    109     SFSIIA P TGVGKT
PkoRG_T_A_Pk     89     SFSIIA P TGMGKS
ApeR2_T_A_Ap    105     SFAIIA P TGVGKS
NeqRG_T_A_Ne     90     SFSIIA P TGMGKT
TpeRG_T_A_Tp    110     SFVILA P TGVGKT
TmaRG_T_B_Tm     90     SFIMVA P TGVGKT
StoR1_T_A_St     93     SFSLSA P TGLGKT
AfuRG_T_A_Af     72     SFAATA P TGVGKT
PaeRG_T_A_Pa     98     SFAIVA P TGSGKT
AaeR2_T_B_Aa    106     SFAIVA P TGVGKT
PabRG_T_A_Pa     81     SFSIIA P TGMGKS
PfuRG_T_A_Pf     88     SFSIIA P TGMGKS
AaeR1_T_B_Aa     86     SFAMLA P TGVGKT
StoR2_T_A_St     91     SFAIIA P PGLGKT
SsoR1_T_A_Ss     92     SFAIIA P PGLGKT
MjaRG_T_A_Mj     95     SFSIVV P TGVGKS
TteRG_T_B_Tt     90     SFTLVA P TGVGKT
SacR2_T_A_Sa     95     SFSLSA P TGVGKT
TthR1_T_B_Tt     81     SFAMLA P TGIGKT
PhoRG_T_A_Ph     81     SFSIIA P TGMGKS
SacR1_T_A_Sa     91     SFAIIA P PGLGKT
```

1. Construction...

ApeR1_T_A_Ap	109	SFSIIA **P** **TGVGKT**
TpeRG_T_A_Tp	110	SFVILA **P** **TGVGKT**
TmaRG_T_B_Tm	90	SFIMVA **P** **TGVGKT**
AfuRG_T_A_Af	72	SFAATA **P** **TGVGKT**
AaeR2_T_B_Aa	106	SFAIVA **P** **TGVGKT**
AaeR1_T_B_Aa	86	SFAMLA **P** **TGVGKT**
TteRG_T_B_Tt	90	SFTLVA **P** **TGVGKT**
SacR2_T_A_Sa	95	SFSLSA **P** **TGVGKT**

NLD : décodage N-local

1. Construction...

```
ApeR1_T_A_Ap    109     SFSIIA P TGVGKT
PkoRG_T_A_Pk     89     SFSIIA P TGMGKS

NeqRG_T_A_Ne     90     SFSIIA P TGMGKT
TpeRG_T_A_Tp    110     SFVILA P TGVGKT
TmaRG_T_B_Tm     90     SFIMVA P TGVGKT

AfuRG_T_A_Af     72     SFAATA P TGVGKT

AaeR2_T_B_Aa    106     SFAIVA P TGVGKT
PabRG_T_A_Pa     81     SFSIIA P TGMGKS
PfuRG_T_A_Pf     88     SFSIIA P TGMGKS
AaeR1_T_B_Aa     86     SFAMLA P TGVGKT

TteRG_T_B_Tt     90     SFTLVA P TGVGKT
SacR2_T_A_Sa     95     SFSLSA P TGVGKT

PhoRG_T_A_Ph     81     SFSIIA P TGMGKS
```

Eduardo Corel, LSG, 2010

NLD : décodage N-local

1. Construction...

```
ApeR1_T_A_Ap    109     SFSIIA P TGVGKT
PkoRG_T_A_Pk     89     SFSIIA P TGMGKS
ApeR2_T_A_Ap    105     SFAIIA P TGVGKS
NeqRG_T_A_Ne     90     SFSIIA P TGMGKT
TpeRG_T_A_Tp    110     SFVILA P TGVGKT
TmaRG_T_B_Tm     90     SFIMVA P TGVGKT

AfuRG_T_A_Af     72     SFAATA P TGVGKT

AaeR2_T_B_Aa    106     SFAIVA P TGVGKT
PabRG_T_A_Pa     81     SFSIIA P TGMGKS
PfuRG_T_A_Pf     88     SFSIIA P TGMGKS
AaeR1_T_B_Aa     86     SFAMLA P TGVGKT


TteRG_T_B_Tt     90     SFTLVA P TGVGKT
SacR2_T_A_Sa     95     SFSLSA P TGVGKT

PhoRG_T_A_Ph     81     SFSIIA P TGMGKS
```

Eduardo Corel, LSG, 2010

1. Exemple de segments regroupés par le NLD sur 23 RG (N=7)

```
ApeR1_T_A_Ap    109     SFSIIA P TGVGKT
PkoRG_T_A_Pk     89     SFSIIA P TGMGKS
ApeR2_T_A_Ap    105     SFAIIA P TGVGKS
NeqRG_T_A_Ne     90     SFSIIA P TGMGKT
TpeRG_T_A_Tp    110     SFVILA P TGVGKT
TmaRG_T_B_Tm     90     SFIMVA P TGVGKT

AfuRG_T_A_Af     72     SFAATA P TGVGKT
PaeRG_T_A_Pa     98     SFAIVA P TGSGKT
AaeR2_T_B_Aa    106     SFAIVA P TGVGKT
PabRG_T_A_Pa     81     SFSIIA P TGMGKS
PfuRG_T_A_Pf     88     SFSIIA P TGMGKS
AaeR1_T_B_Aa     86     SFAMLA P TGVGKT
StoR2_T_A_St     91     SFAIIA P PGLGKT
SsoR1_T_A_Ss     92     SFAIIA P PGLGKT

TteRG_T_B_Tt     90     SFTLVA P TGVGKT
SacR2_T_A_Sa     95     SFSLSA P TGVGKT

PhoRG_T_A_Ph     81     SFSIIA P TGMGKS
SacR1_T_A_Sa     91     SFAIIA P PGLGKT
```

Eduardo Corel, LSG, 2010

1. Exemple de segments regroupés par le NLD sur 23 RG (N=7)

```
ApeR1_T_A_Ap    109     SFSIIA P TGVGKT
PkoRG_T_A_Pk     89     SFSIIA P TGMGKS
ApeR2_T_A_Ap    105     SFAIIA P TGVGKS
NeqRG_T_A_Ne     90     SFSIIA P TGMGKT
TpeRG_T_A_Tp    110     SFVILA P TGVGKT
TmaRG_T_B_Tm     90     SFIMVA P TGVGKT
StoR1_T_A_St     93     SFSLSA P TGLGKT
AfuRG_T_A_Af     72     SFAATA P TGVGKT
PaeRG_T_A_Pa     98     SFAIVA P TGSGKT
AaeR2_T_B_Aa    106     SFAIVA P TGVGKT
PabRG_T_A_Pa     81     SFSIIA P TGMGKS
PfuRG_T_A_Pf     88     SFSIIA P TGMGKS

StoR2_T_A_St     91     SFAIIA P PGLGKT
SsoR1_T_A_Ss     92     SFAIIA P PGLGKT

TteRG_T_B_Tt     90     SFTLVA P TGVGKT
SacR2_T_A_Sa     95     SFSLSA P TGVGKT

PhoRG_T_A_Ph     81     SFSIIA P TGMGKS
SacR1_T_A_Sa     91     SFAIIA P PGLGKT
```

Eduardo Corel, LSG, 2010

1. Exemple de segments regroupés par le NLD sur 23 RG (N=7)

```
SsoR2_T_A_Ss      99     SFIMSA P TGLGKT
ApeR1_T_A_Ap     109     SFSIIA P TGVGKT
PkoRG_T_A_Pk      89     SFSIIA P TGMGKS
ApeR2_T_A_Ap     105     SFAIIA P TGVGKS
NeqRG_T_A_Ne      90     SFSIIA P TGMGKT
TpeRG_T_A_Tp     110     SFVILA P TGVGKT
TmaRG_T_B_Tm      90     SFIMVA P TGVGKT
StoR1_T_A_St      93     SFSLSA P TGLGKT
AfuRG_T_A_Af      72     SFAATA P TGVGKT
PaeRG_T_A_Pa      98     SFAIVA P TGSGKT
AaeR2_T_B_Aa     106     SFAIVA P TGVGKT
PabRG_T_A_Pa      81     SFSIIA P TGMGKS
PfuRG_T_A_Pf      88     SFSIIA P TGMGKS
AaeR1_T_B_Aa      86     SFAMLA P TGVGKT
StoR2_T_A_St      91     SFAIIA P PGLGKT
SsoR1_T_A_Ss      92     SFAIIA P PGLGKT

TteRG_T_B_Tt      90     SFTLVA P TGVGKT
SacR2_T_A_Sa      95     SFSLSA P TGVGKT

PhoRG_T_A_Ph      81     SFSIIA P TGMGKS
SacR1_T_A_Sa      91     SFAIIA P PGLGKT
```

Eduardo Corel, LSG, 2010

NLD : décodage N-local

1. Exemple de segments regroupés par le NLD sur 23 RG (N=7)

SsoR2_T_A_Ss	99	SFIMSA **P** TGLGKT
ApeR1_T_A_Ap	109	SFSIIA **P** TGVGKT
PkoRG_T_A_Pk	89	SFSIIA **P** TGMGKS
ApeR2_T_A_Ap	105	SFAIIA **P** **TGVGKS**
NeqRG_T_A_Ne	90	SFSIIA **P** TGMGKT
TpeRG_T_A_Tp	110	SFVILA **P** TGVGKT
TmaRG_T_B_Tm	90	SFIMVA **P** TGVGKT
StoR1_T_A_St	93	SFSLSA **P** TGLGKT
AfuRG_T_A_Af	72	SFAATA **P** TGVGKT
PaeRG_T_A_Pa	98	SFAIVA **P** TGSGKT
AaeR2_T_B_Aa	106	SFAIVA **P** TGVGKT
PabRG_T_A_Pa	81	SFSIIA **P** TGMGKS
PfuRG_T_A_Pf	88	SFSIIA **P** TGMGKS
AaeR1_T_B_Aa	86	**SFAMLA** **P** **TG**VGKT
StoR2_T_A_St	91	SFAIIA **P** PGLGKT
SsoR1_T_A_Ss	92	SFAIIA **P** PGLGKT
MjaRG_T_A_Mj	95	SFSIVV **P** **TGVGKS**
TteRG_T_B_Tt	90	SFTLVA **P** TGVGKT
SacR2_T_A_Sa	95	SFSLSA **P** TGVGKT
TthR1_T_B_Tt	81	**SFAMLA** **P** **TG**IGKT
PhoRG_T_A_Ph	81	SFSIIA **P** TGMGKS
SacR1_T_A_Sa	91	SFAIIA **P** PGLGKT

Eduardo Corel, LSG, 2010

Taille de l'environnement de décodage?

- Dans le dernière implémentation de la méthode N s'adapte automatiquement à la séquence en cours d'analyse selon le contexte local (VLD : Variable Length local Decoding).

- VLD : généralisation de la méthode de décodage.

- Pour en savoir plus:

Didier G., Corel E., Laprevotte I., Grossmann A. & Landès-Devauchelle C. (2012). "Variable length local decoding and alignment-free sequence comparison". *Theoretical Computer Science*, **462** : 1-11.

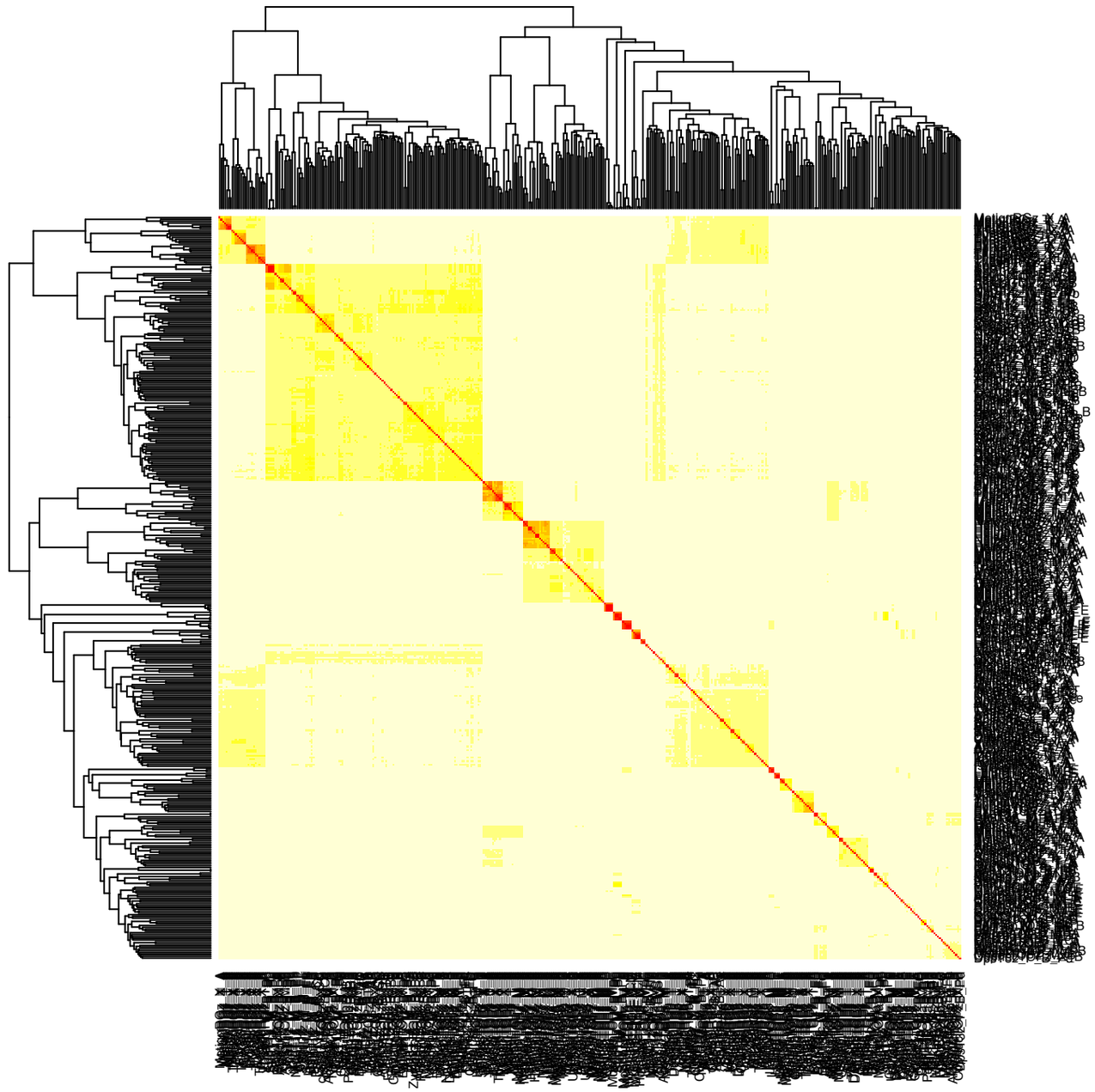
Calcul d'une matrice de distances

- La similarité entre deux séquences est le nombre de caractères décodés communs divisé par la taille de la séquence la plus courte.

- La dissimilarité est égale à 1 - la similarité.

- On calcule la matrice de dissimilarité pour tous les couples de séquences du jeu de données.

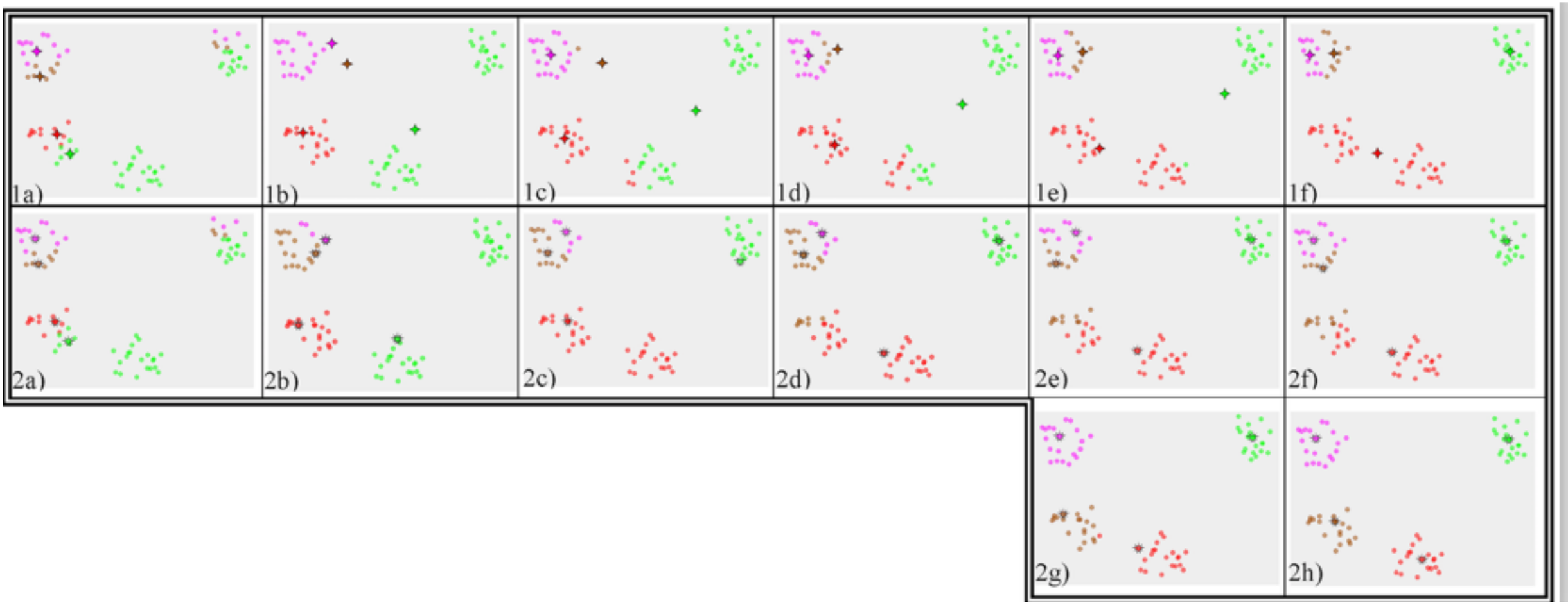
Quel clustering sur la matrice de dist?



Quel clustering sur la matrice de dist?

- Automatiser procédure classification en distinguant les groupes forts des groupes faibles
- Quelle stratégie pour faire le clustering?
- Travail d'Antoine Garnier (M1 informatique UA)

Partitions around medoids (pam)



Répartir en combien de groupes?

k-medoids selon wikipedia (k-medoid vs k-means)

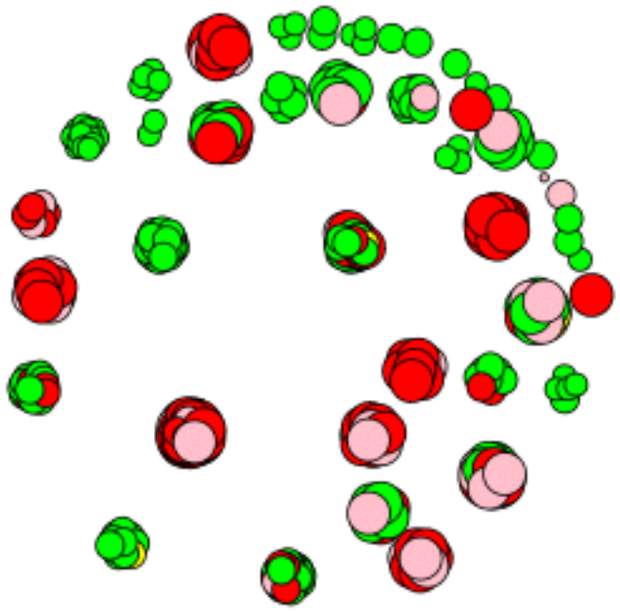
Recherche de formes fortes

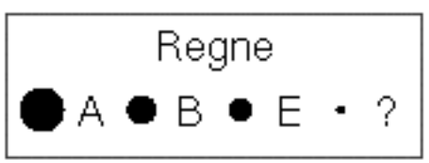
- Faire des tentatives pour plusieurs répartition en k groupes (ici de k=5 à k = 25) soit 21 tentatives

- Mettre ensemble les individus qui ont toujours été regroupés ensemble selon les 21 tentatives

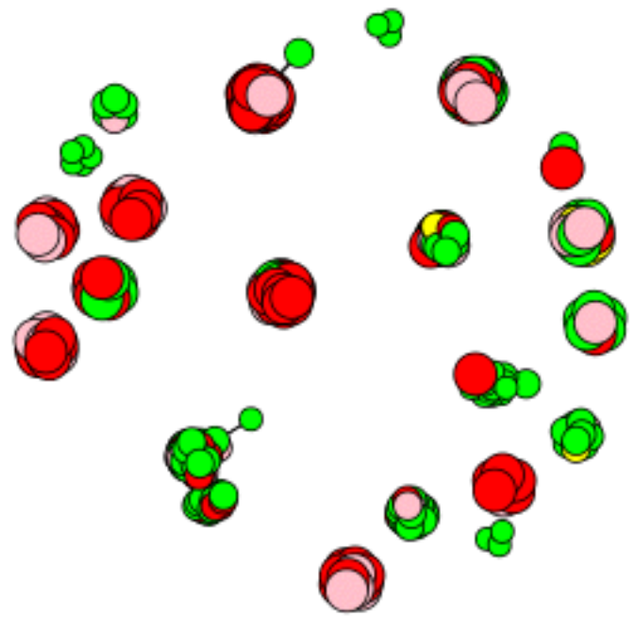
- Ce sont des formes fortes (tjs ou svt classés ens)

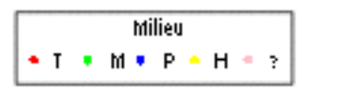
Recherche de formes fortes

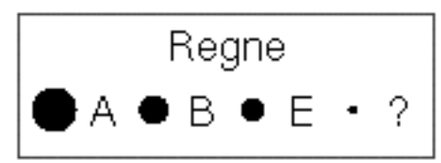




21 fois ensemble

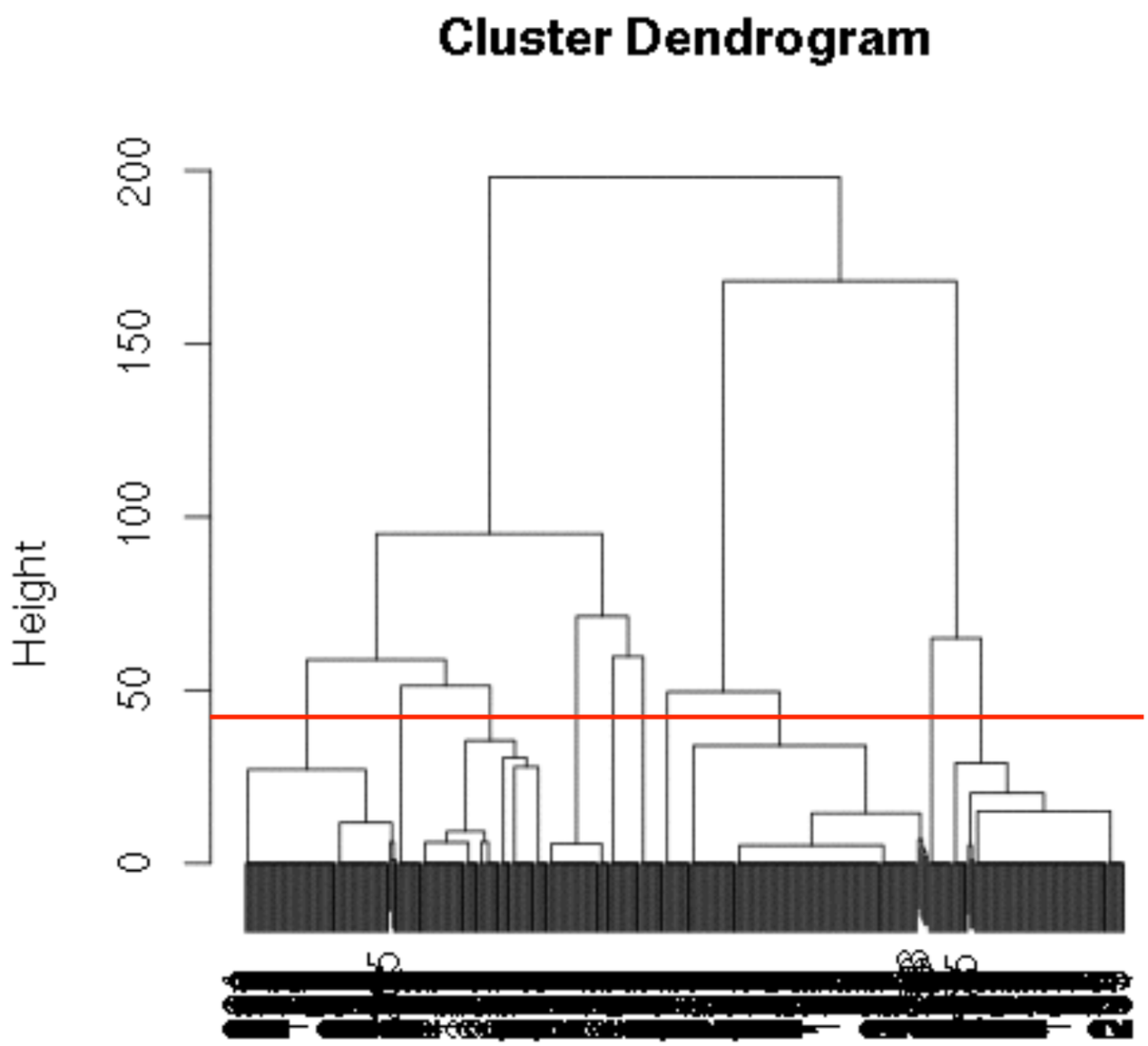






16 ou + fois ensemble

Clustering hiérarchique (hclust)



cuttree (10 grp)

Résultats

```
Nombre de groupes :  10

##########################################
####
# Groupe  1  -  79  proteines
reverse gyrase (7 intrus)
##########################################
####
# Groupe  2  -  109  proteines
topoisomerases bacteries
##########################################
####
# Groupe  3  -  28  proteines
topoisomerases archees (1 intrus)
##########################################
####
# Groupe  4  -  70  proteines
« bacteries » (42 B, 25 A, 3 E Leishmania)
##########################################
####
# Groupe  5  -  12  proteines
topoisomerases bacteries
##########################################
####
# Groupe  6  -  58  proteines
« eucaryotes » (31 E, 19 A, 8B)
##########################################
####
# Groupe  7  -  14  proteines
archees (halophiles)
##########################################
####
# Groupe  8  -  11  proteines
archees (methanogenes)
##########################################
####
# Groupe  9  -  11  proteines
RG archees (pyrococcales)
##########################################
####
# Groupe  10  -  11  proteines
top archees(pyrococcales)
```

```
Nombre de groupes :  16

##########################################
72 reverse gyrase (7 intrus) (G1)
 7 reverse gyrase            (G14)
##########################################
21 topoisomerases bacteries (G2)
88 topoisomerases bacteries (G4)
##########################################
28 topoisomerases archees (1 intrus) (G3)

##############################################
(# ex Groupe  4  -  70  proteines)
groupe 5 : 42 topoisomerases bacteries
groupe 6 : 28 topoisomerases archees (+3E)
##############################################

##########################################
12 topoisomerases bacteries (G7)
##########################################

##############################################
(# ex Groupe  6  -  58  proteines)
groupe 8 : 26 eucaryotes + 8 archees (36)
groupe 9 : 11 archees                (11)
groupe 10 : 5 eucaryotes             ( 5)
groupe 11 : 6 bacteries              ( 6)
##############################################

14 topoisomerases archees (halophiles)(G12)
##########################################
11 topoisomerases archees (methanogenes) (G13)
##########################################
RG archees (pyrococcales) (G15)
##########################################
11 topoisomerases archees(pyrococcales) (G16)
```

Conclusion

- Résultats préliminaires valident la méthode des formes fortes
- Expertises en cours pour fixer la granulosité de la classification
- Classification en cours sur les dernières extraction des topoisomérases dans les génomes
- Objectif à cours terme : mettre à jour TopoIBase pour publication (expertise classification) - outil d'aide à la notation des topoisomérases de type IA
- Recherche expert des topoisomérases humaines (eucaryotes)

Perspectives

- TopoIBase : base de données experte de topoisomérases de type IA dotée de plusieurs outils (http://stat.genopole.cnrs.fr/topodb)

- Annotations enrichies par les experts du domaine (forum modéré)

- A moyen-long terme incorporation des topo IB et des topo II

- Travail présenté à workshop EMBO 2015 par **Marc Nadal**
  EMBO Workshop : DNA topoisomerases, DNA topology and human health, 13 – 17 September 2015 | Les Diablerets, Switzerland

- Travail présenté à JOBIM 2015 par **Nicolas Daccord**

Remerciements

Implementation de TopoIBase

Nicolas Daccord (1-4), Damien Correia (1)

Franck Samson (1) & Vladimir Daric (2)

Expertises des Topoisomérases et Filtres de Détection

Florence Vogliolo(1), Damien Correia (1), Anaïs Louis (1)

Gilles Grasseau (1), Hélène Debat (2-6) & Marc Nadal (2-6)

Classification sans alignements

Anaïs Louis (1), Claudine Landès-Devauchelle (1-4),

Antoine Garnier (4), Eduardo Corel (1-5) & Gilles Didier (3)

1 - Laboratoire de Mathématique et Modélisation d'Evry (Evry)

2- Institut de Génétique Microbienne (Orsay)

3 - Laboratoire d'Informatique de Luminy (Marseille)

4- Institut de Recherche en Horticulture et Semences (Angers)

5- Institut de Biologie Paris Seine (Paris)

6- Institut Jacques Monot (Paris)