Génomique comparée et génomique fonctionnelle à Bordeaux

David Sherman

3ème rencontres autour de
la plateforme Bio-informatique

Rennes
2005-10-18

Plateforme Génomique Fonctionnelle

Groupe Loisier Architecte mandataire - Raguenaud et Roux Architecte associé

- Bio-informatique
- Protéomique
- Transcriptome
- Génotypage-séquençage
- Microscopie électronique
- Microscopie photonique

Centre de Bio-informatique de Bordeaux

- Responsable : A. de Daruvar
- Plate-forme opérationnelle bio-informatique (enquête RIO 2003)
- Moyens importants de stockage et de calcul
- Prestations : 4 ingénieurs
  - Mise à disposition de ressources, développements logiciels à façon, analyse de données, formations
- Recherche : 10 chercheurs et enseignants-chercheurs U.B2, U.B1/ENSEIRB (LaBRI)

LaBRI UMR 5800

- Directeur : S. Dulucq
- Fédère (presque) toute la recherche en Informatique des universités de la région
- 250 personnes (130 chercheurs/ens-chercheurs)
- 5 équipes dont INRIA Futurs
- Projet transversal AMBI – Méthodes et algorithmes pour la Bio-informatique (nouvelle équipe à compter de 2006)

Objectifs – Construction de *vues globales* et *multi-échelles* du vivant

- Génomique comparée
  - Génolevures — levures hémiascomycètes
  - Molligen — mollicutes
  - Oenigène — oenococcus oenii
- IntAct & HUPO PSI — réseaux d'interactions
- BlastSets — intégration de données
- S. siliceus — modélisation discrète
- Arborescences — architecture de plantes

Génomique comparée

- Génolevures – levures
  - http://cbi.labri.fr/Genolevures
- Molligen – mollicutes
  - Large spectre pathogène : animaux, plantes, homme
  - Petits génomes
  - http://cbi.labri.fr/molligen
- Oenigène – *O. oeni*
  - Intérêt oenologique : fermentation malo-lactique
  - Annotation et comparison de gènes

Interactions protéine-protéine

- Au coeur de la plupart des **processus**
  - Transduction de signal
  - Voies métaboliques
  - Réponse immunitaire **...**
- Éléments primitifs du **comportement**
- **Approches globales** expérimentales
  - Petite échelle, grande échelle, haut débit
- Construction de **réseaux d'interaction**

Le projet européen « IntAct »

Une banque de données fédérée d'interactions protéine-protéine

- Banque de données et outils à source ouverte
- Système de noeuds locaux
  - Protection de données avant publication
  - Échanges entre noeuds
- Interface d'interrogation Web
- Outils d'analyse et de visualisation
- Définition de **vocabulaires contrôlés**
- Standards internationaux
  - HUPO PSI mi, ms

[Hermjakob *et al*, **NAR**, 2004]

[Hermjakob *et al*, **Nature Biotech**, 2004]

Visualisation avec ProViz

- Identification de protéines ou d'interactions
  - Par mot-clé ou par analyse du graphe
- Comparaison visuelle de graphes
- Extraction de vues, sous-graphes d'intérêt
  - Chaque vue a ses propres paramètres
  - Organisé dans un graphe quotient
- Hautement interactif (grâce à Tulip) [Auber, 2003]
- Analyses, clustering (grâce aux plug-ins)

[Iragne *et al*, **Bioinformatics**, 2004]

Views | Protein Ontology | Interaction CV | Properties

[+/-] All

Unselect All

Find

Next

Description	go
Gene_Ontology	GO:0003673
biological_process	GO:0008150
behavior	GO:0007610
biological_process unknown	GO:0000004
cell communication	GO:0007154
cell adhesion	GO:0007155
cell-cell signaling	GO:0007267
cell invasion	GO:0030260
cell recognition	GO:0008037
host-pathogen interaction	GO:0030383
pollen-pistil interaction	GO:0009875
response to endogenous stimulus	GO:0009719
response to external stimulus	GO:0009605
signal transduction	GO:0007165
cell growth and/or maintenance	GO:0008151
death	GO:0016265
development	GO:0007275
obsolete	GO:0008371
physiological processes	GO:0007582
viral life cycle	GO:0016032
cellular_component	GO:0005575
cell	GO:0005623
cellular_component unknown	GO:0008372
external protective structure	GO:0030312
extracellular	GO:0005576
immunoglobulin complex	GO:0019814
obsolete	GO:0008370
unlocalized	GO:0005941
virion	GO:0019012
molecular_function	GO:0003674
anticoagulant	GO:0008435
antifreeze	GO:0016172
antioxidant	GO:0016209
apoptosis regulator	GO:0016329
binding	GO:0005488
cell adhesion molecule	GO:0005194
chaperone	GO:0003754
chaperone regulator	GO:0030188
cytoskeletal regulator	GO:0008580

Interaction CV | Properties

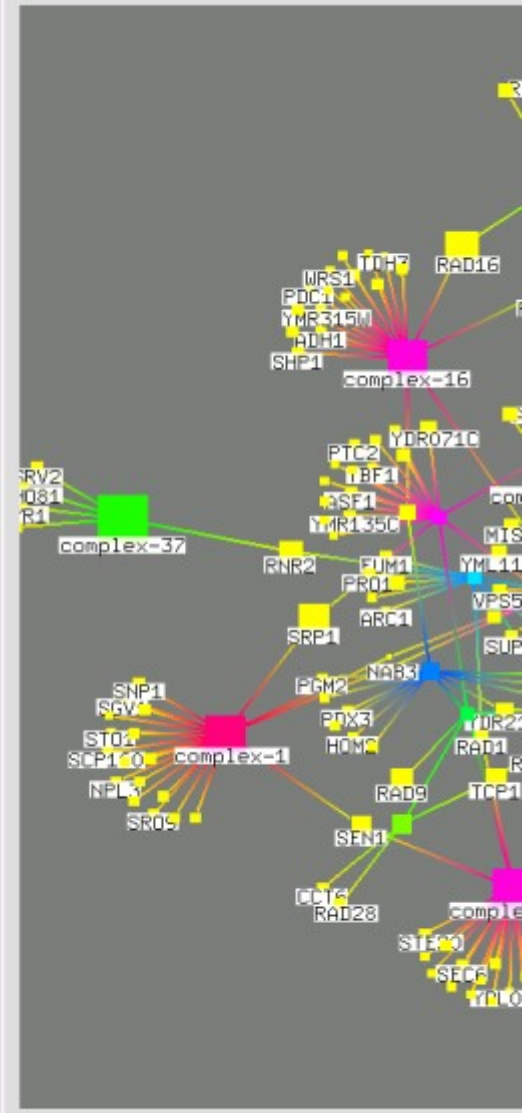
Select All | Next

Selected

Node 252

	Value
	(10,1.9,1)
	|GO:0000108|GO:0000113|GO:
	Nucleotide excision repair prot
	|GO:0000108|GO:0000113|GO:
e	IDA|TAS
etection_CV	
	RAD16
	(211,-190.5,0)
	false
	Cone
	(18.3848,18.3848,18.3848)

root



Clone view | Select All

Close view | Deselect All

Extraction dictée par une politique

Interactions

HUPO PSI

Voies métaboliques

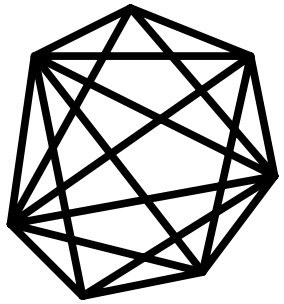
BioPAX

Protéine-ligande

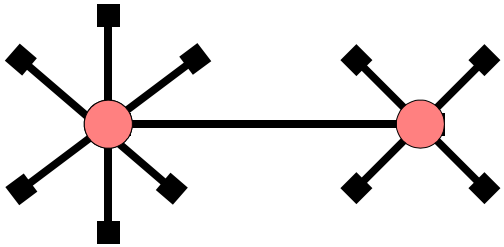
divers

États et isoformes

Uniprot & divers



Graphe binaire
-- partenaires



Graphe « *spoke* »
-- complexes

Résumés complexes

Synthèse de réseaux consensuels

Modèles stochastiques

[Iragne, **en cours**]

BlastSets

- Intégration de données produites à l'échelle cellulaire par des procédés post-génomiques
  - Génomique fonctionnelle
  - Rapprochement dynamique d'informations hétérogènes
- Comprendre l'articulation entre les mécanismes moléculaires et les fonctions cellulaires
- Approche par *classifications BlastSets* et une *recherche efficace* d'ensembles cibles

BlastSets

[Barriot *et al*, **NAR**, 2004]

BlastSets - Mozilla

http://cbi.labri.fr/outils/BlastSets/

QUERY | TASK LIST | CREDITS | MANUAL

start session

BlastSets

STEP 1 Species: Saccharomyces cerevisiae

STEP 2 Query set(s): Sequence IDs / Classification

yil169c
yhr214w-a
yar068w
yol155c

Or Keywords - SWISSPROT_Keywords

STEP 3 Target sets:

Systematic classification - ENZYME
Systematic classification - GO - biological process
Systematic classification - GO - cellular component
Systematic classification - GO - molecular function
Systematic classification - KEGG SCe
Systematic classification - SUBCELL

Expression data | ADD | REMOVE | CLEAR | INVERT

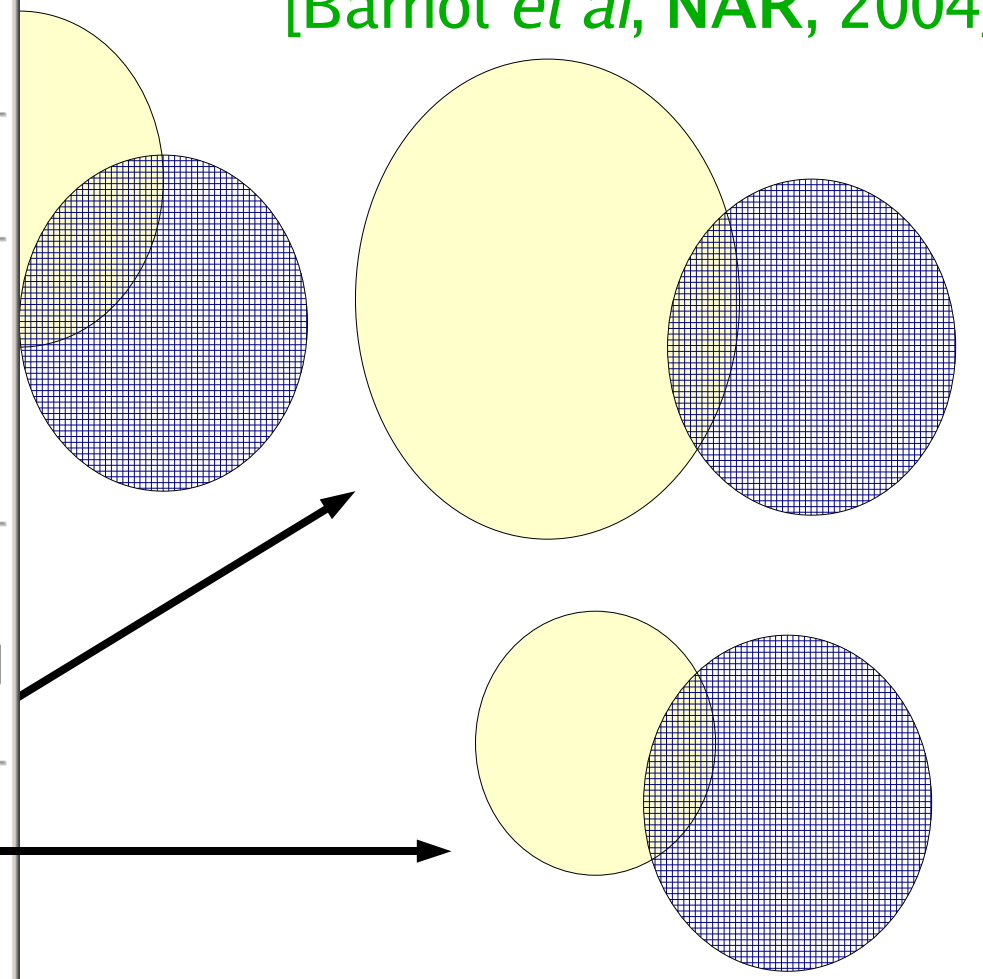
STEP 4 Results filters

Evalue: Use a cut off on Evalue 0.1

Results: Keep all results / Only 20 results

STEP 5 Query name: *(optional)*

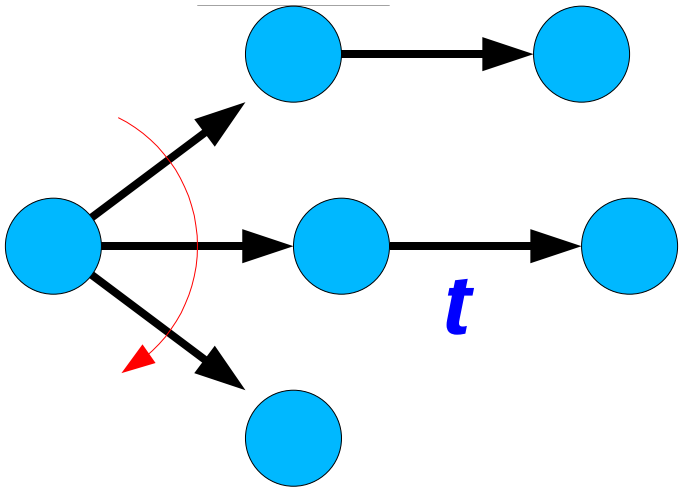
SUBMIT



Mesure de similitude probabalistique

Saccharomyces siliceus

- Les modèles du comportement cellulaire doivent tenir compte de phénomènes discrétes (M.Aigle)
- Développement d'outils de modélisation
  - **Discrètes**
  - **Multi-échelle**
  - **Stochastiques**



Transitions probabalistes

Transitions temporisés

[Soueidan, Nikolski 2005; YSBN]

Arborescences

- Données organisées en arborescences
- Analyse de la variabilité d'un ensemble d'individus
  - Amélioration des algorithmes globaux
  - Prise en compte d'infos biologiques structurales
  - Contraintes multi-échelles
- Analyse structurelle d'un individu
  - Étiquetage automatique
  - Identification de motifs et autosimilarité

Approches multi-échelles

- Génomique comparée
  & interactions

  organisation de génome
  évolution moléculaire
  réseaux dynamiques
  extraction par politique

- BlastSets

  échelle de la cellule
  classifications structurées

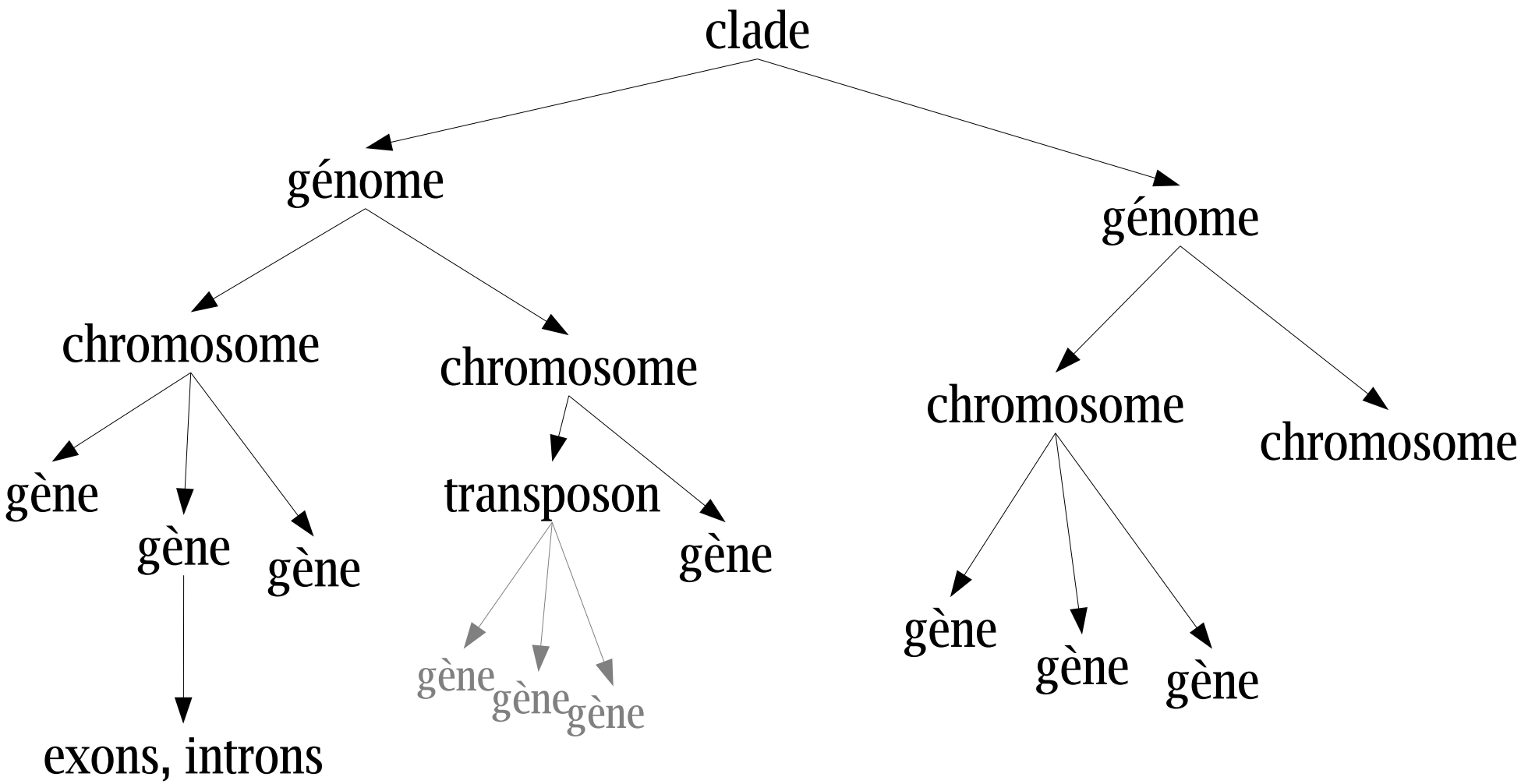
- *S. siliceus*

  modèles discrètes, multi-échelles et stochastiques du comportement

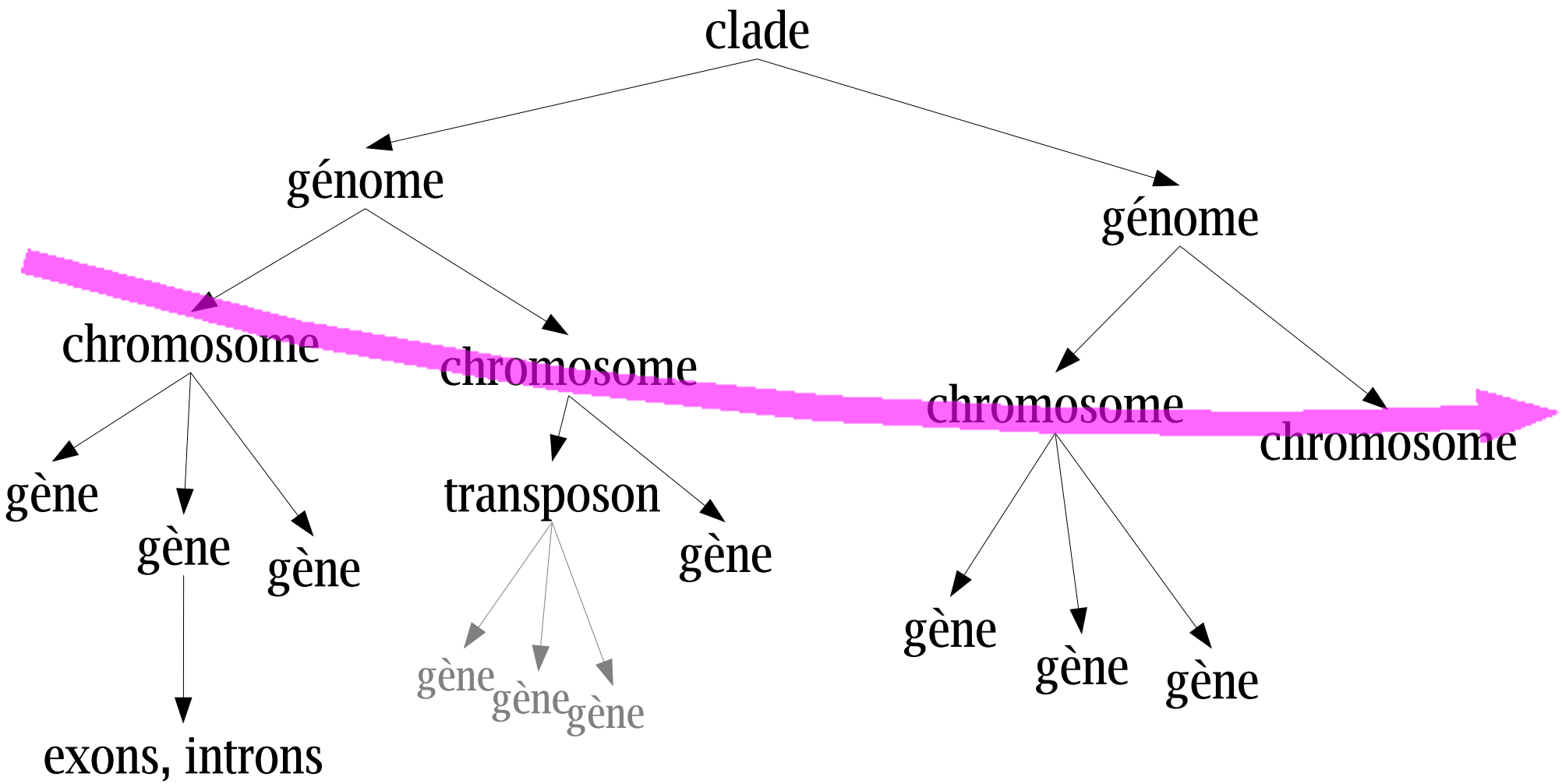
- Arborescences

  niveaux de description, propriétés autosimilares

Échelles dans la génomique



Échelles dans la génomique

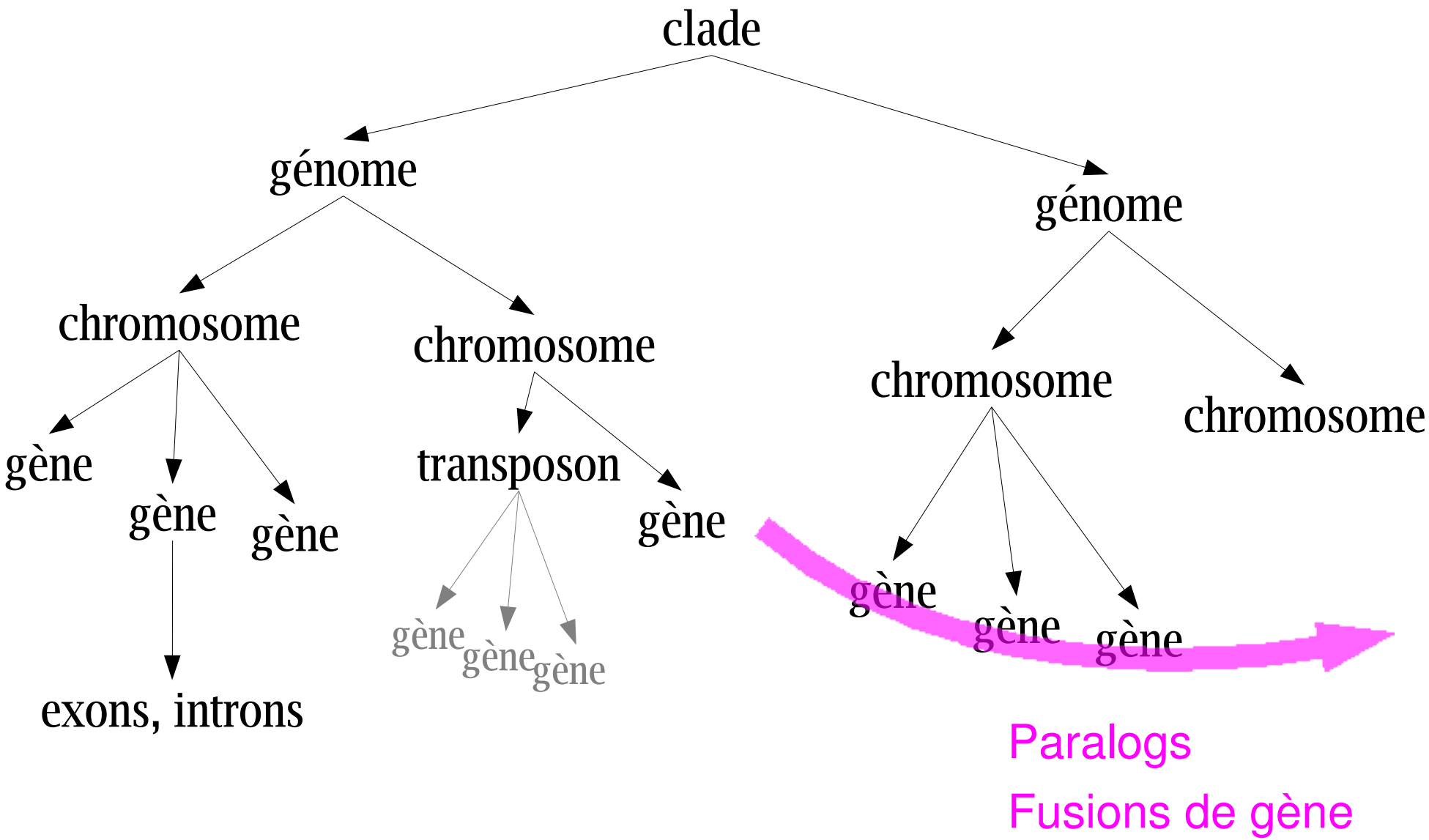


Duplication segmentale

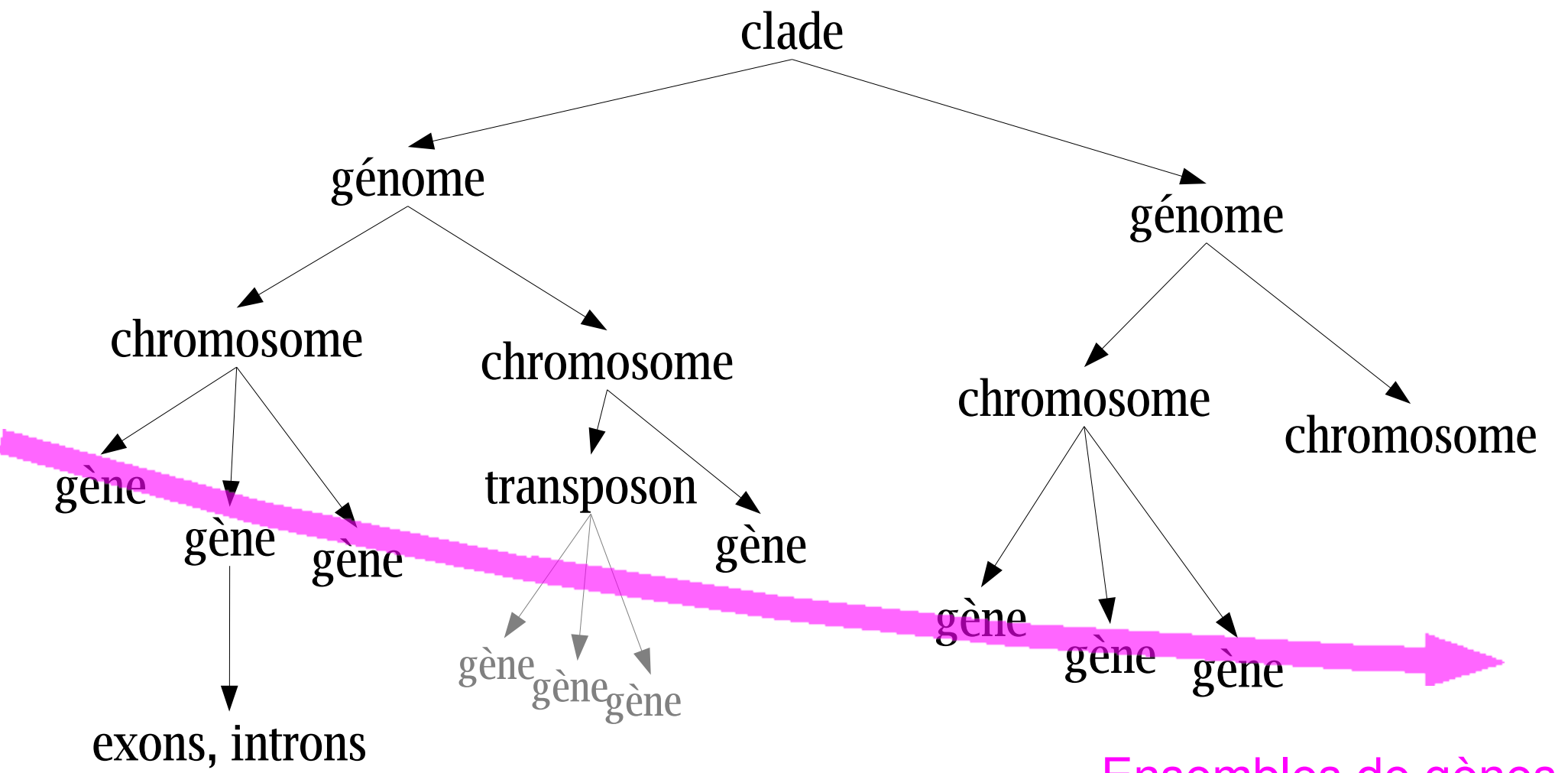
Translocations

WGD

Échelles dans la génomique



Échelles dans la génomique

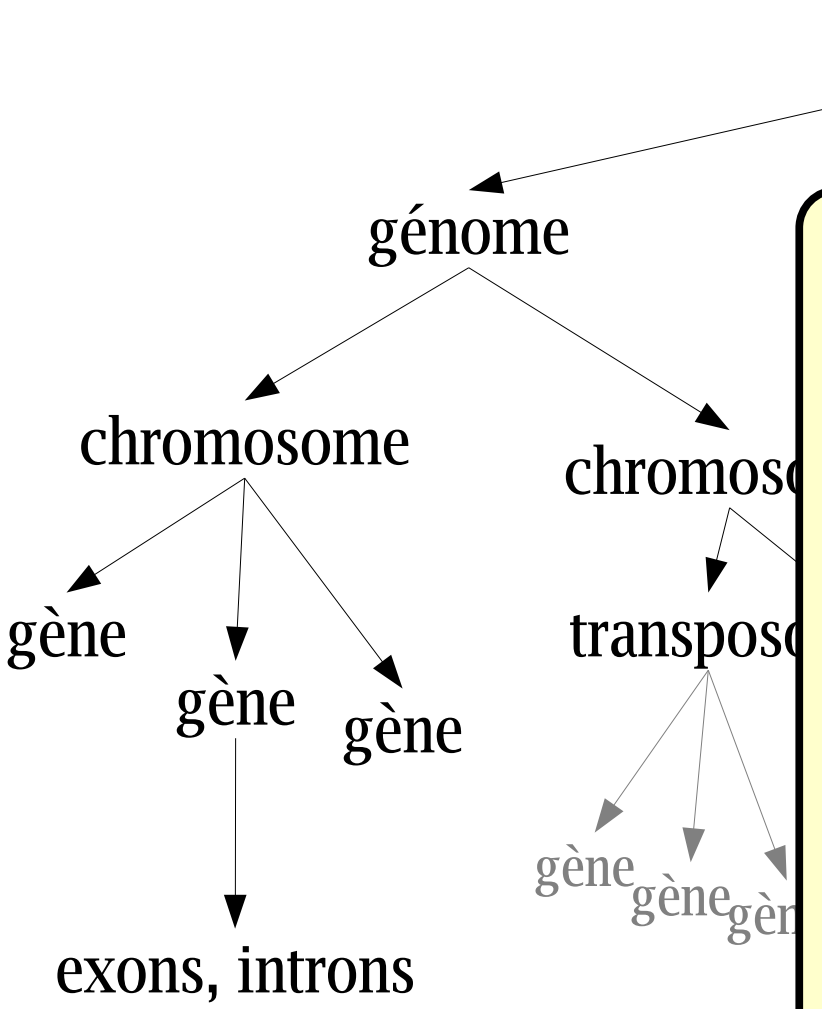


Ensembles de gènes

Familles de protéines

Synteny

Échelles dans la génomique



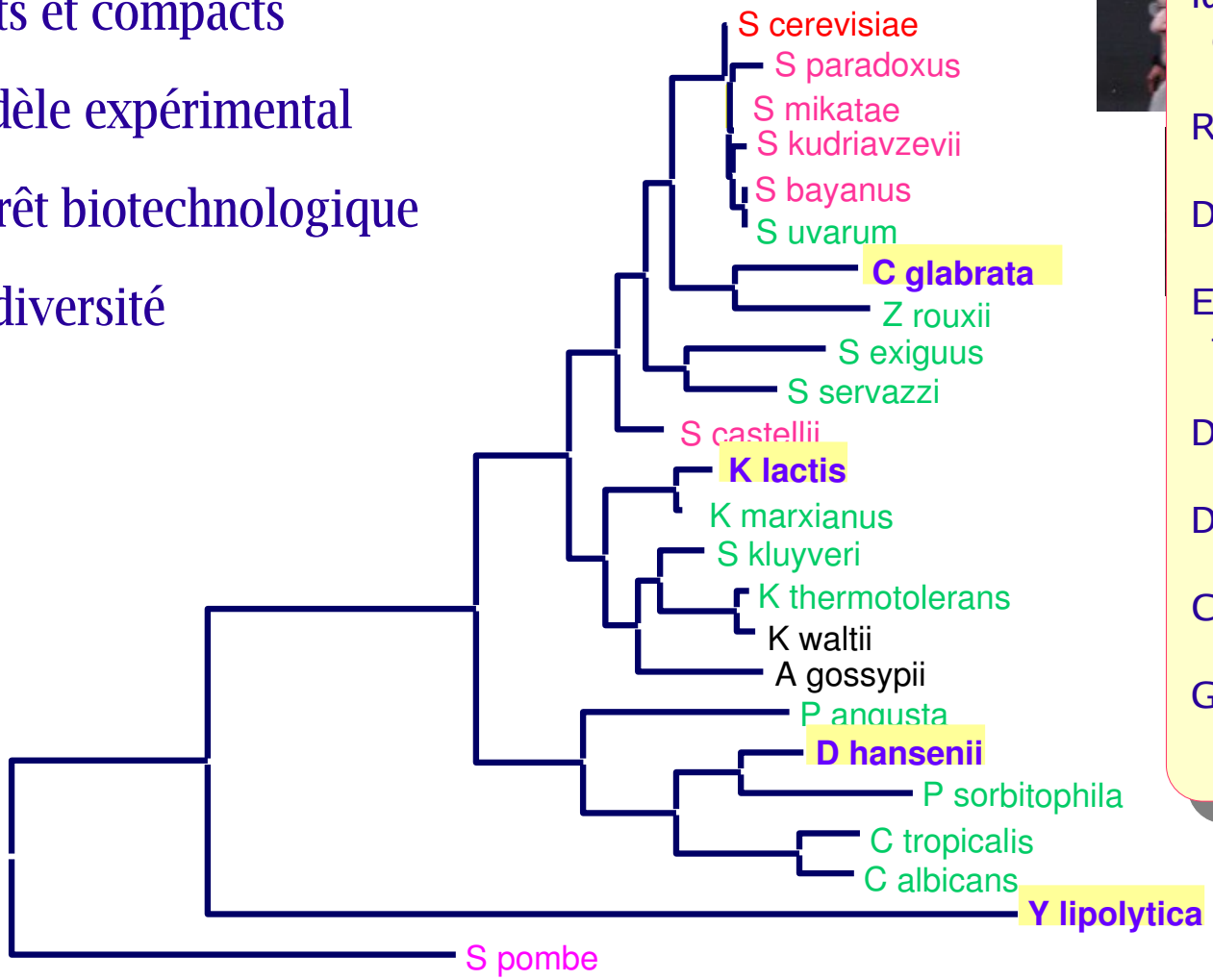
Structure de génomes

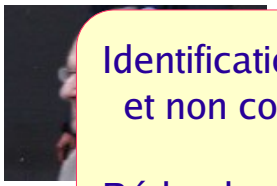
- Évolution moléculaire
  - événements passés
  - mécanismes distincts
- Comportement cellulaire
  - voies métaboliques
  - réseaux d'interaction
  - modèles

Le programme « Génolevures »

Comparaison à grande échelle de levures hémiascomycètes

- Petits et compacts
- Modèle expérimental
- Intérêt biotechnologique
- Biodiversité





Identification de séq. codantes et non codantes

Rédondance des génomes

Divergence ortho-/para- logues

Expansion et contraction de familles universelles

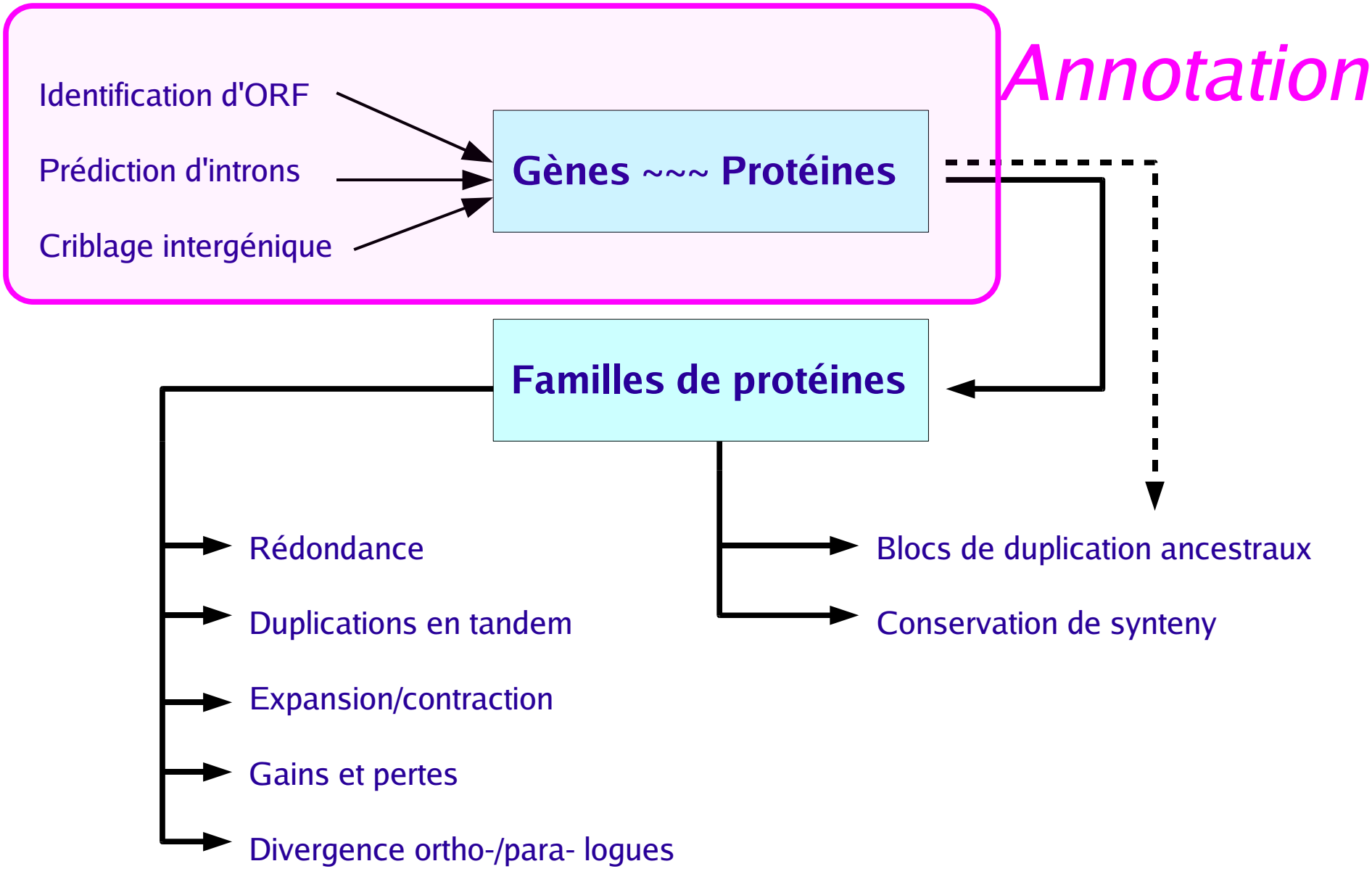
Duplications en tandem

Dups. de blocs ancestraux

Conservation de synteny

Gain et perte de gènes

Méthodes et analyses



Page d'anno-tation

Decision

Infos

Browser

Navigation

Analyses

[Frangeul *et al*, 2004]

File Edit View Go Bookmarks Tools Window Help

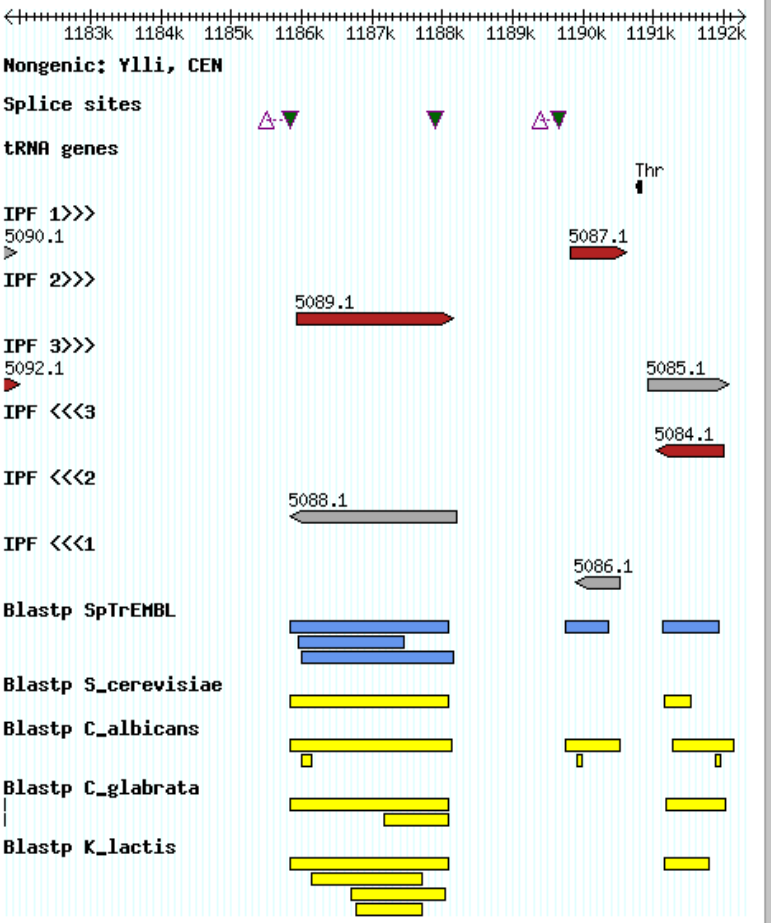
Back Forward Reload Stop http://cbi.labri.fr/Genolevures/Y_lipolytica/cgi-bin/IPF_reader.CGI?IPF_name=5 Search Print

5092.1<-previous (3939 b) IPFutil - all_IPF (1644 b) next->5087.1

5089.1

Highly similar to sp|P19414 Saccharomyces cerevisiae Aconitate hydratase, mitochondrial precursor (EC 4.2.1.3) (Citrate hydro-lyase) (Aconitase)

length = 747 AA
Yali0D from 1185923 to 1188166 (p)
ORF sequence comes from 1185842 to 1188166
wide nucleotide sequence comes from 1185342 to 1188366
GC% = 56 3rd=56



Links to Curator tasks:

- Decide whether True ORF or FALSE ORF
  - GeneMark signal?
  - Overlaps?
  - Blastp, Tblastx?
- Identify START codon
- Define a coding sequence
- Assign a putative function

Genome browser focus on IPF 5089.1
Genome Browser **Fasta dump** of IPF 5089.1 and 1000bp upstream

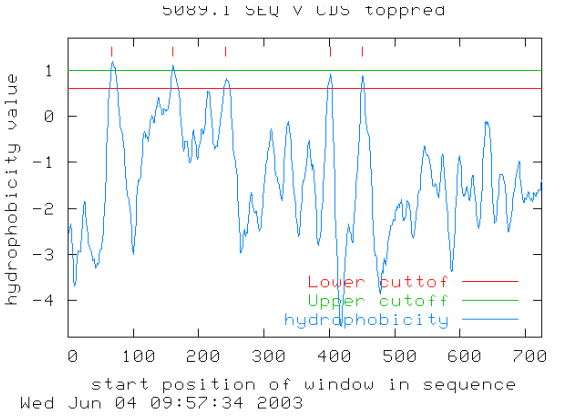
BEST-BLASTP:

```
>Neurospora_crassa_3.aa 79%    Identities = 519/746 (69%), Positives = 595/746 (79%), Gaps = 1/746 (0%

    wi|NCU02366.1 Neurospora crassa NCU02366.1 hypothetical protein          (63698 - 66657)
>SpTrEMBL 88%    Identities = 590/747 (78%), Positives = 660/747 (88%), Gaps = 1/747 (0%)

    sp|P19414 Saccharomyces cerevisiae Aconitate hydratase, mitochondrial precursor (EC 4.2.1.3) (Cit
>S_cerevisiae.aa 87%    Identities = 582/747 (77%), Positives = 650/747 (87%), Gaps = 13/747 (1%)

    YLR304c  sp|P19414 Saccharomyces cerevisiae YLR304c ACO1 aconitate          hydratase P4 9 f4 1
```

Analyses pour l'expert

V_COMMENT: Change_V_COMMENT

```
Intron predicted.
>Intron1 354nt
GTGAGtatcgatggtgaagaaagacaccgacaatcgccacgttgtgccacagacacagacgcgtt
>EXON1
ATGCTGGCTTCTCGAGTTTCCATCAAGGCT
```

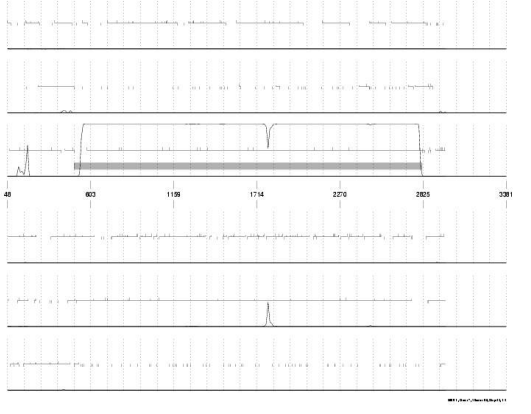


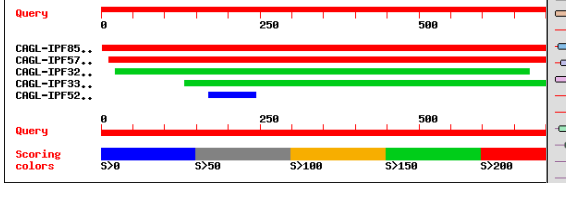
USER COMMENT :

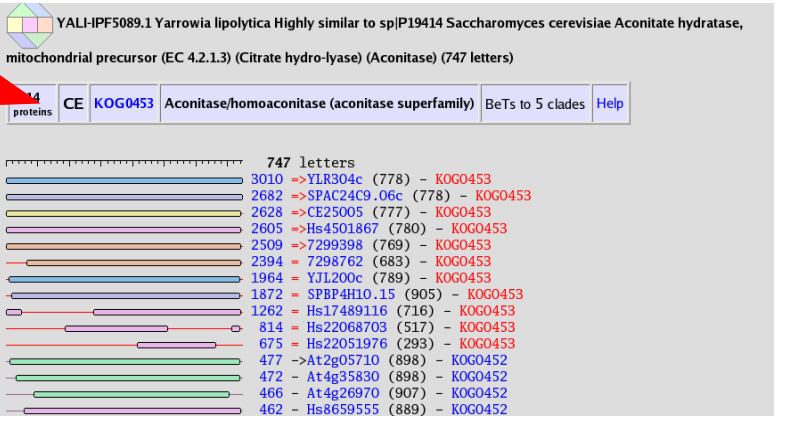
= Empty =

IPF Field				
DOMAIN_NUCL	ALIAS_CDS	V_COMMENT	WHOSWHO	V_EXON
BEST-BLASTP	BDBH	V_NOTE	ADMIN_BATCH	V_RBS
MW-IP	SEQCDS	COMMENT_SEQCDS	SEQNUCLARGE	SEQNUCL
SEQAA				

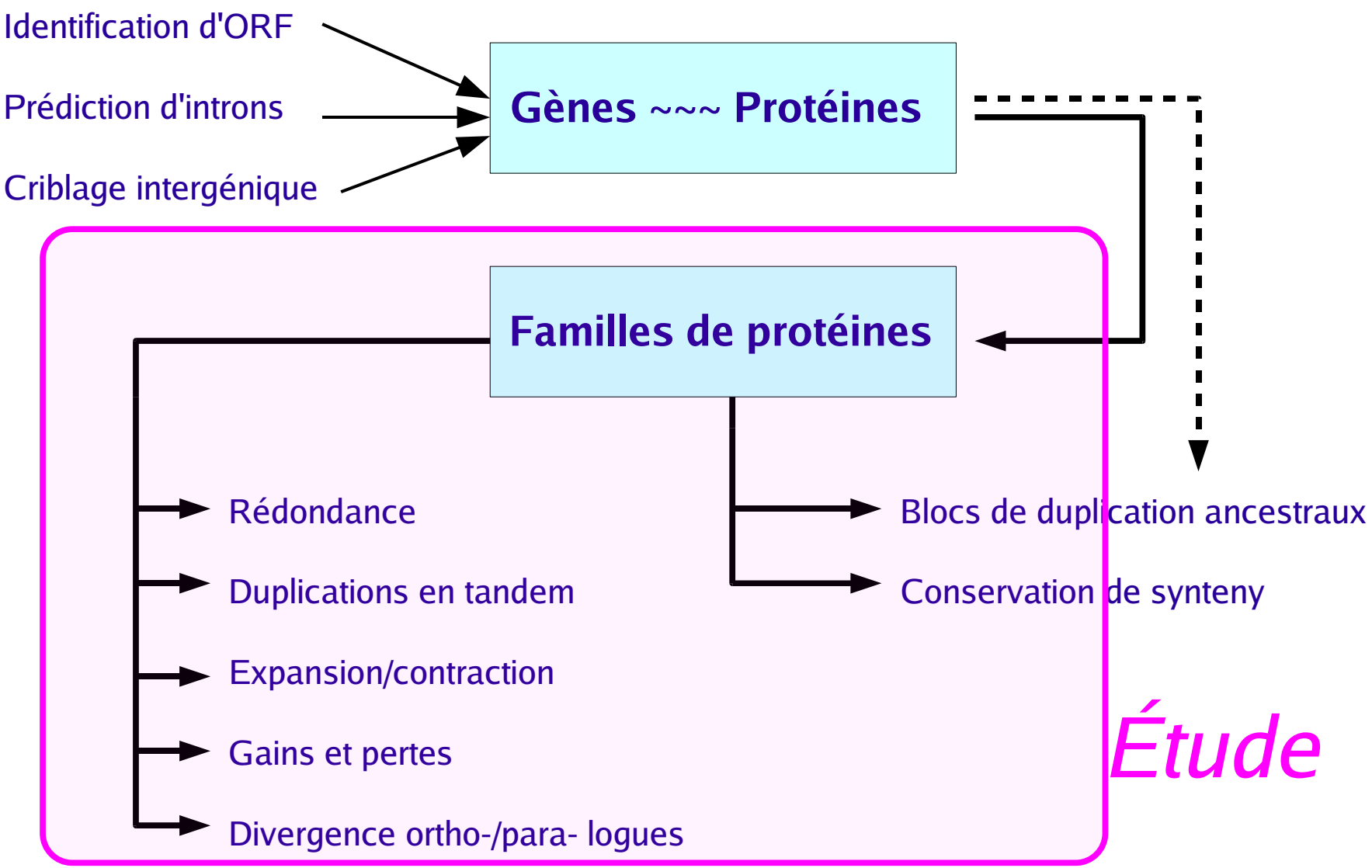
IPF Results			
toppred SEQ_V_CDS	toppred SEQAA	Align B20	Best-GL1 alt
Best-GL1 std	AUTO blastp	D_hansenii blastp	K_lactis blastp
C_glabrata blastp	S_cerevisiae blastp	C_albicans blastp	SpTrEMBL blastp
N_crassa blastp	Euk COG	prev gm graph	prev gm txt
gm graph ORF=500 on frame 3	gm txt ORF=500 on frame 3		



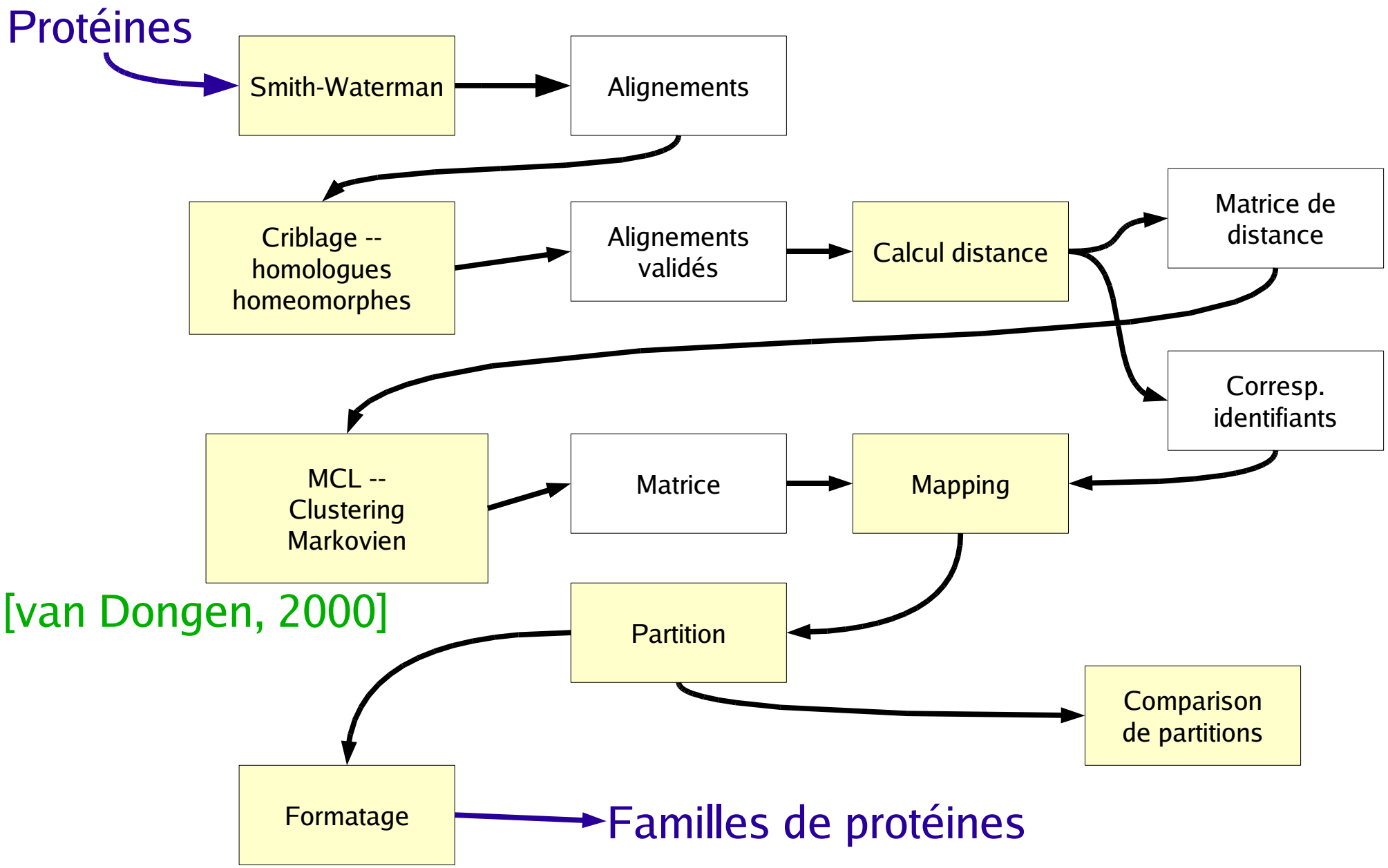




Méthodes et analyses



Calcul de familles de protéines



Comparaison de patterns phylétics

pattern	Families nb	Proteins nb
Family universal to the 5 yeasts		
sckdy	2,014	17,153
Families restricted to *S. cerevisiae*, *C. glabrat*		
sck--	572	1,827
sc---	245	547
s-k--	102	212
-ck--	34	77
Family restricted to *D. hansenii* and *Y. lipolyt*		
---dy	488	1,121
Families missing in one species only		
-ckdy	17	100
s-kdy	49	312
sc-dy	44	215
sck-y	81	382
sckd-	316	1,445
Species-specific families		
s----	31	95
-c---	8	51
--k--	16	38
---d-	135	501
----y	177	492
All other combinations		
sc-d-	16	53

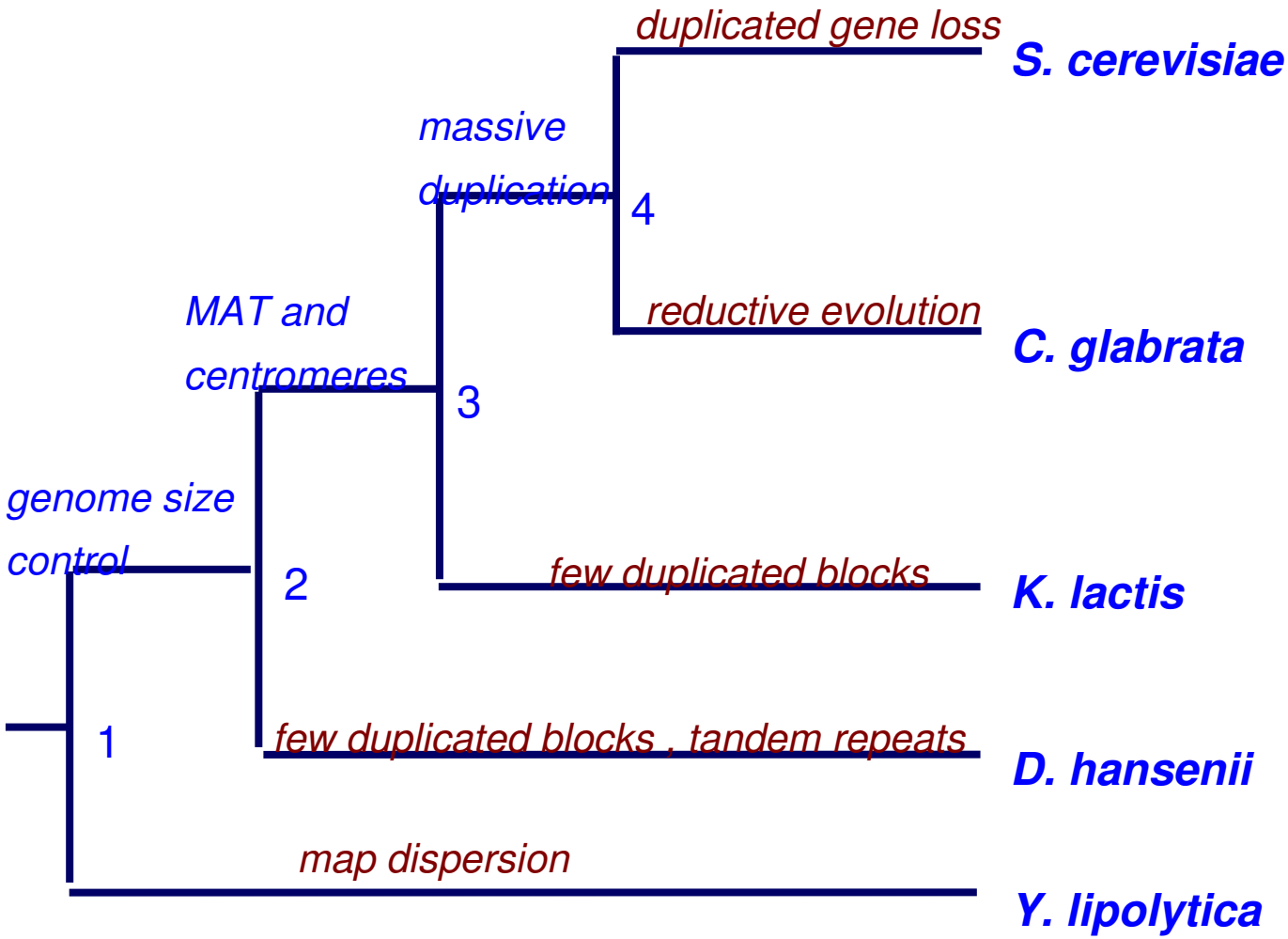
s · c · k · d · y

Presence dans l'espèce corresp.

Effectifs dans l'espèce corresp.

```
Family      Members      Pattern  Profile

GLR.230:  5-member family  sckdy  (1 1 1 1 1)
GLR.3:   30-member family  sckdy  (6 6 6 6 6)
GLR.208:  6-member family  sckdy  (2 1 1 1 1)
GLR.21:  13-member family  sckdy  (2 2 2 4 3)

GLR.335:  5-member family  sck--  (2 2 1 0 0)
GLR.356:  5-member family  sck--  (2 2 1 0 0)
GLR.383:  4-member family  sck--  (1 1 2 0 0)

GLR.994:  5-member family  s-kdy  (2 0 1 1 1)
GLR.1212: 4-member family  s-kdy  (1 0 1 1 1)
GLR.1219: 4-member family  s-kdy  (1 0 1 1 1)

GLR.1220: 4-member family  ---d-  (0 0 0 4 0)
GLR.1550: 2-member family  ---d-  (0 0 0 2 0)
GLR.1551: 2-member family  ---d-  (0 0 0 2 0)
```

L'histoire des génomes de levures

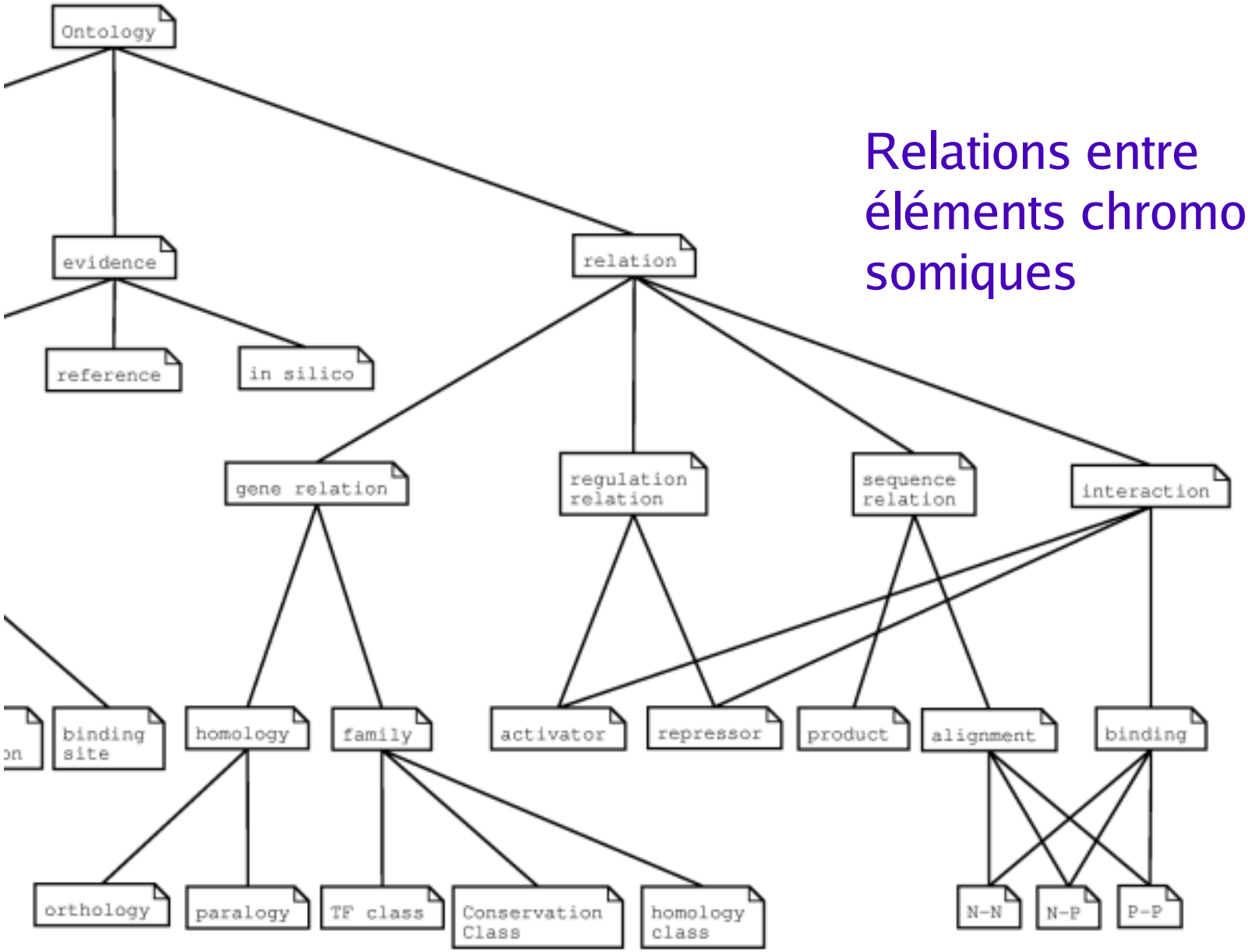


[Dujon, Sherman, Fischer *et al*, **Nature**, 2004]

Vocabulaire contrôlé

Relations entre éléments chromo-somiques

(SOFA)



[Sherman, Durrens *et al*, **NAR**, 2004]

Génolevures:

Genomic Exploration of the Hemiascomycete Yeasts

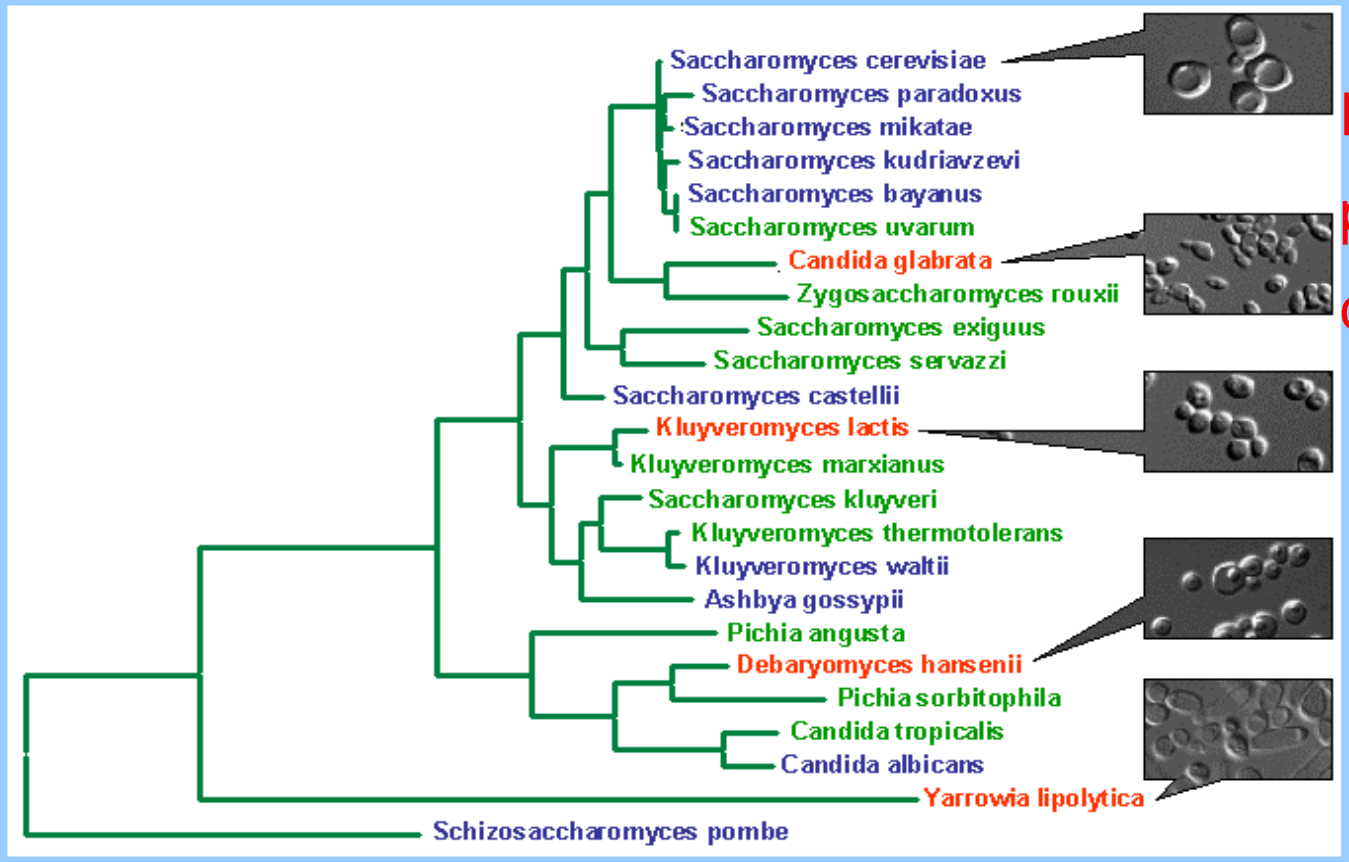
Help

Search

Tools

- About Génolevures
- The Consortium
- Help!
- Nomenclature
- Search Génolevures
- Protein Families
- Blast
- Download
- Publications
- Related links

Génolevures is a large-scale comparative genomics project across the evolutionary range of the Hemiascomycetous yeast phylum. This site provides tools and data relative to 4 complete genome sequences and 10 partial genome sequences determined by the Génolevures Consortium, and links to other complete or partial yeast genome sequences.



25S phylogeny using maximum parsimony

Links to project overviews

Center for Bioinformatics · Bordeaux
mail Webmaster@cbi.labri.fr

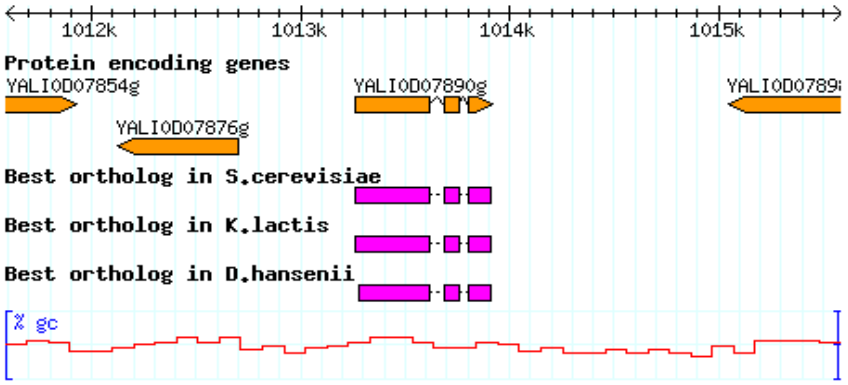
URL: http://cbi.labri.fr/Genolevures/index-site-2.php
Update: Monday, 31 May 2004

Page d'élément chromosomique

YALI0D07890g
highly similar to tr|Q8X0I8 Neurospora crassa Probable E2 ubiquitin-conjugating enzyme, start by similarity

Yarrowia lipolytica Complete CDS.
ORF from Yali0D:1013261..1013919 sense (p)
Predicted protein is 179 aa

Synonyms: YALI0D07890g|YALI-IPF42007



Carte cliquable du voisinage chromosomique

Theoretical

Molecular weight: 20.27127 kDa
Avg residue weight: 113.247 Da
Isoelectric point: 8.0814
Charge: 3.5

(InterPro domains cartoon)

Analyses *In silico*

Protein family

Family GLR.1954 sckdy: 56 proteins (13 10 10 12 12)

Appartenance en famille

Results

S.cerevisiae blastp	Uniprot blastp
C.albicans blastp	N.crassa blastp
C.glabrata blastp	K.lactis blastp
D.hansenii blastp	Y.lipolytica blastp

Séquence décorée

CDS translation product

```
>gnl|GLV|YALI0D07890g highly similar to tr|Q8X0I8 Neurospora crassa Probable
 E2 ubiquitin-conjugating enzyme, start by similarity [Yarrowia lipolytica]
 Complete CDS. Translation product.
MLKIWSMKEKQAAEGGSAKKKVTAAQLRVQKDLSELSLPSTMKTHFPSAEDIMNFELTVR
```

Exploration d'annotations

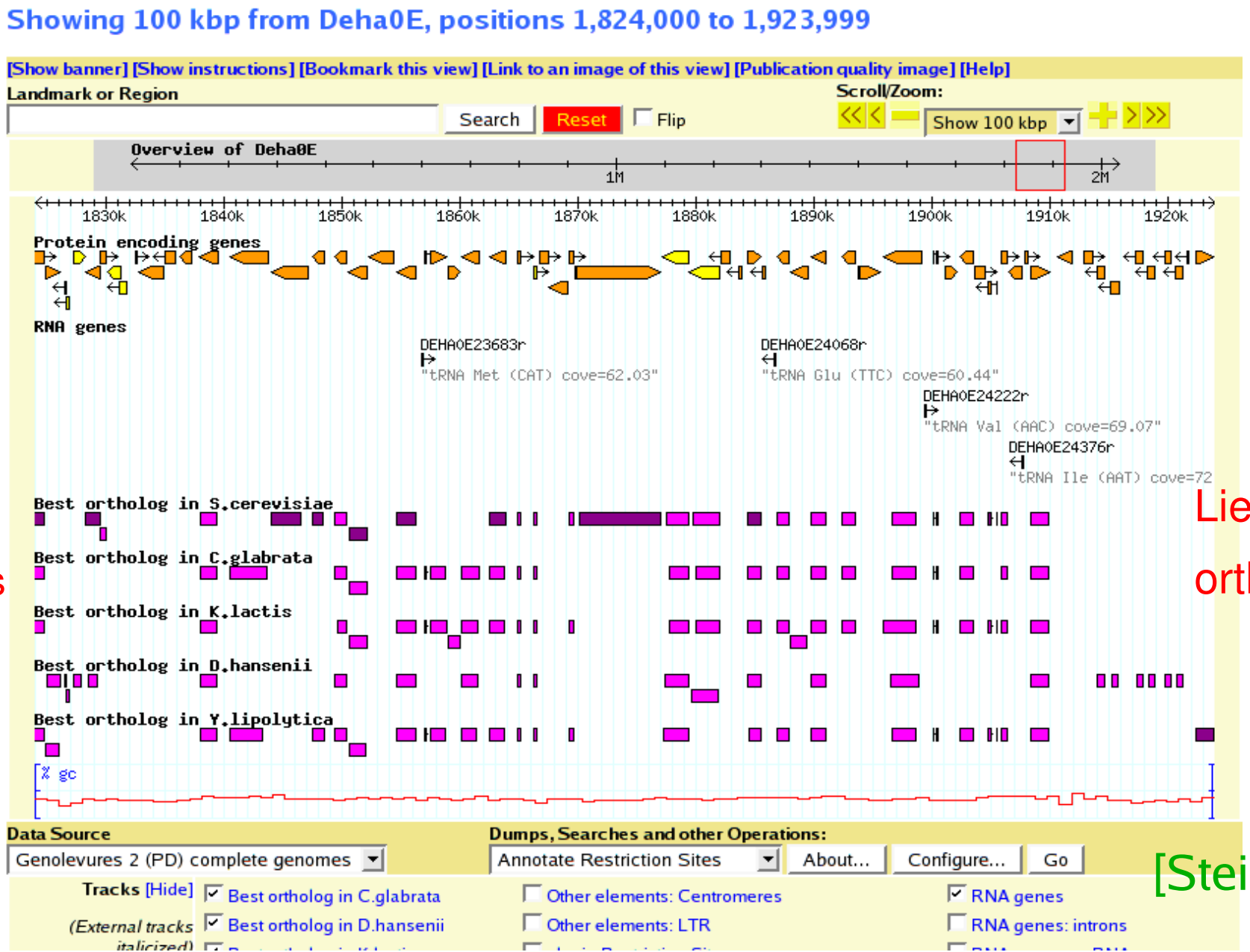
Gènes

RNA,

LTR

Curated

relations



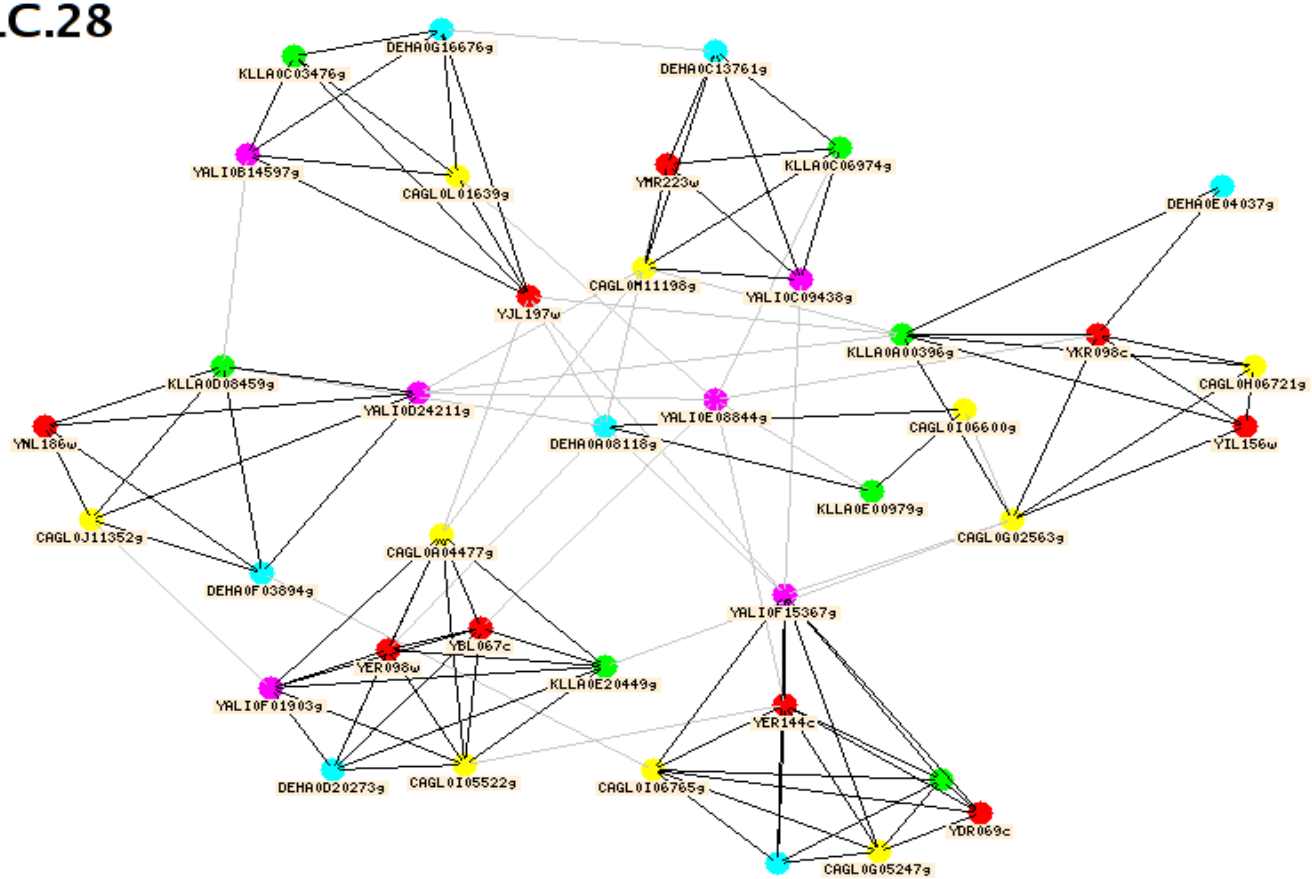
Liens vers

orthologues

[Stein, 2002]

Familles : carte de relations

GLC.28



39 members :

Gene Name	aa length	Defline
YBL067c	688	UBP13 ubiquitin carboxyl-terminal hydrolase
YDR069c	926	DOA4 ubiquitin-specific protease
YER098w	754	UBP9 ubiquitin carboxyl-terminal hydrolase
YER144c	805	UBP5 ubiquitin-specific protease
YIL156w	1071	UBP7 ubiquitin carboxy terminal hydrolase
YJL197w	1254	UBP12 ubiquitin C-terminal hydrolase
YKR098c	717	UBP11 ubiquitin C-terminal hydrolase
YMR223w	471	UBP8 similarity to mouse deubiquitinating enzyme and UBP13P,UBP9,DOA4P
YNL186w	792	UBP10 derepression of telomeric silencing

Points communs

- Production de données
- Annotation primaire
- Analyses et algorithmes
- Intégration et visualisation
- Élaboration de modèles

L'équipe et remerciements

- AMBI au LaBRI
  - Roland Barriot
  - Emmanuelle Beyne
  - Isabelle Dutour
  - Serge Dulucq
  - Pascal Durrens
  - Pascal Ferraro
  - Hélène Ferry
  - Claire Herrbach
  - Florian Iragne
  - Macha Nikolski
  - Aïda Ouangraoua
  - Hayssam Soueidan
- Génolevures
  - Bernard Dujon
  - Jean-Luc Souciet
- IntAct
  - Henning Hermjakob
  - Samuel Kerrien
- CBiB
  - Antoine de Daruvar
  - Elisabeth Bon
  - Alexis Groppi
  - Daniel Jacob

Remerciements

- GDR Génolevures
- ACI IMPBIO
  - BlastSets
  - GL en Ligne
- ACI NIM
- FEDER
- FP5
- FP6

CONSEIL REGIONAL AQUITAINE

Université Victor Segalen Bordeaux 2

UNIVERSITÉ BORDEAUX 1 Sciences Technologies

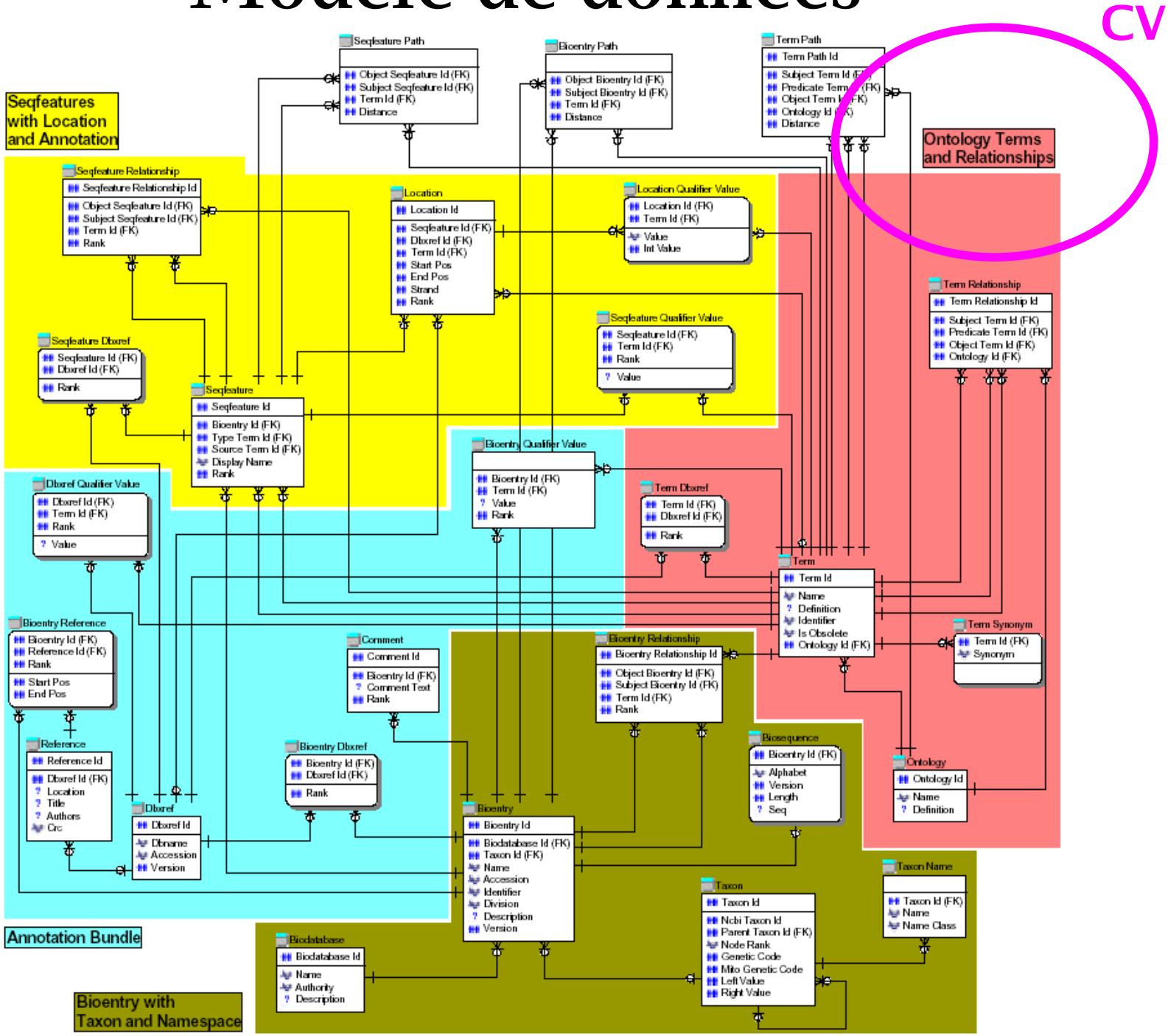
CNRS

INRA

C·H·U Hôpitaux de Bordeaux

Inserm

Modèle de données



[Birney *et al*, 2003]