



Models and algorithms for the genome

2011-01-05

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

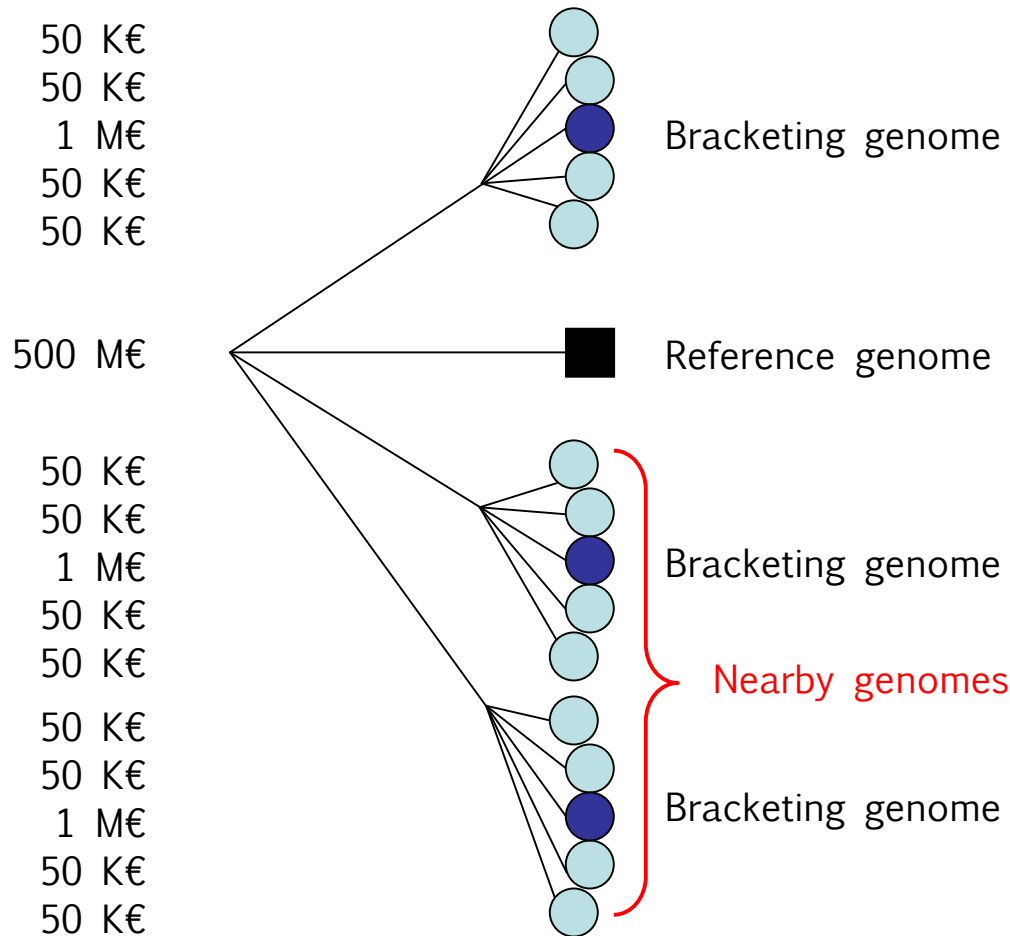


centre de recherche
BORDEAUX - SUD-OUEST

David J. Sherman, DR INRIA

INRIA project-team « MAGNOME »
joint with CNRS & U. Bordeaux

Comparative genomics: a growing challenge



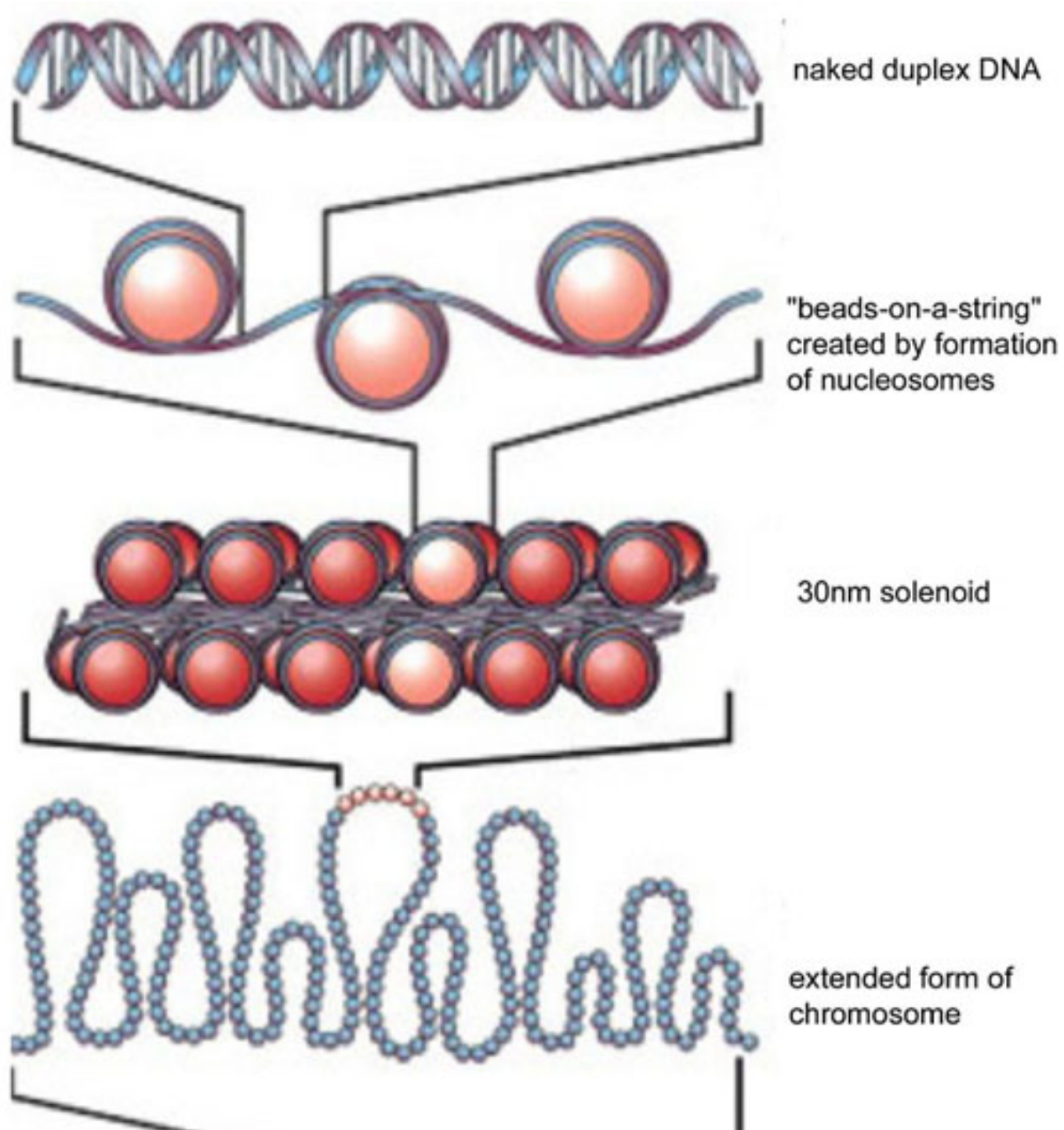
Characterized genes
Validated models

Comparative genomics makes most sense in this vicinity

Very many applications

- Health
- Biotechnologies
- Biodiversity

Essentially scaling challenges for computer science



Landmark or Region

0_721:664816..704815

Search

Reset

Flip

Scroll/Zoom:



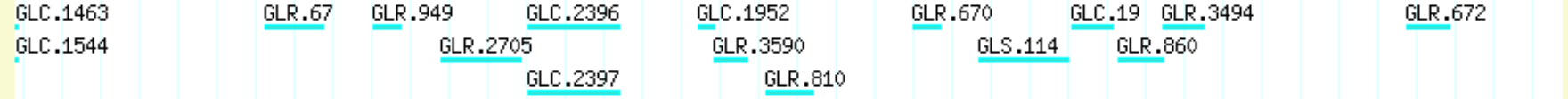
Show 40 kbp



Overview of 0_721



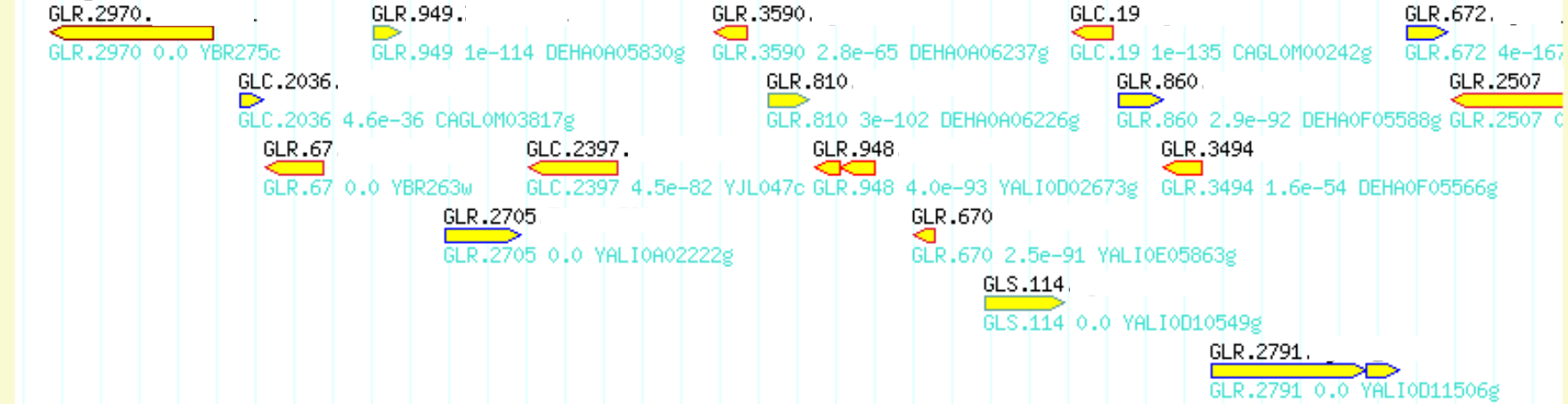
Honology group



IPF (6-frame)



Family PSItblastn



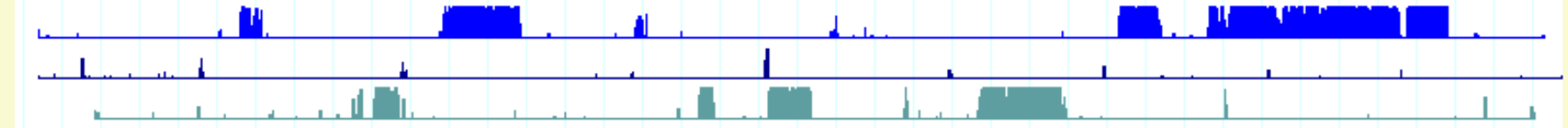
CAAT-Box start codons



GeneMark ROI



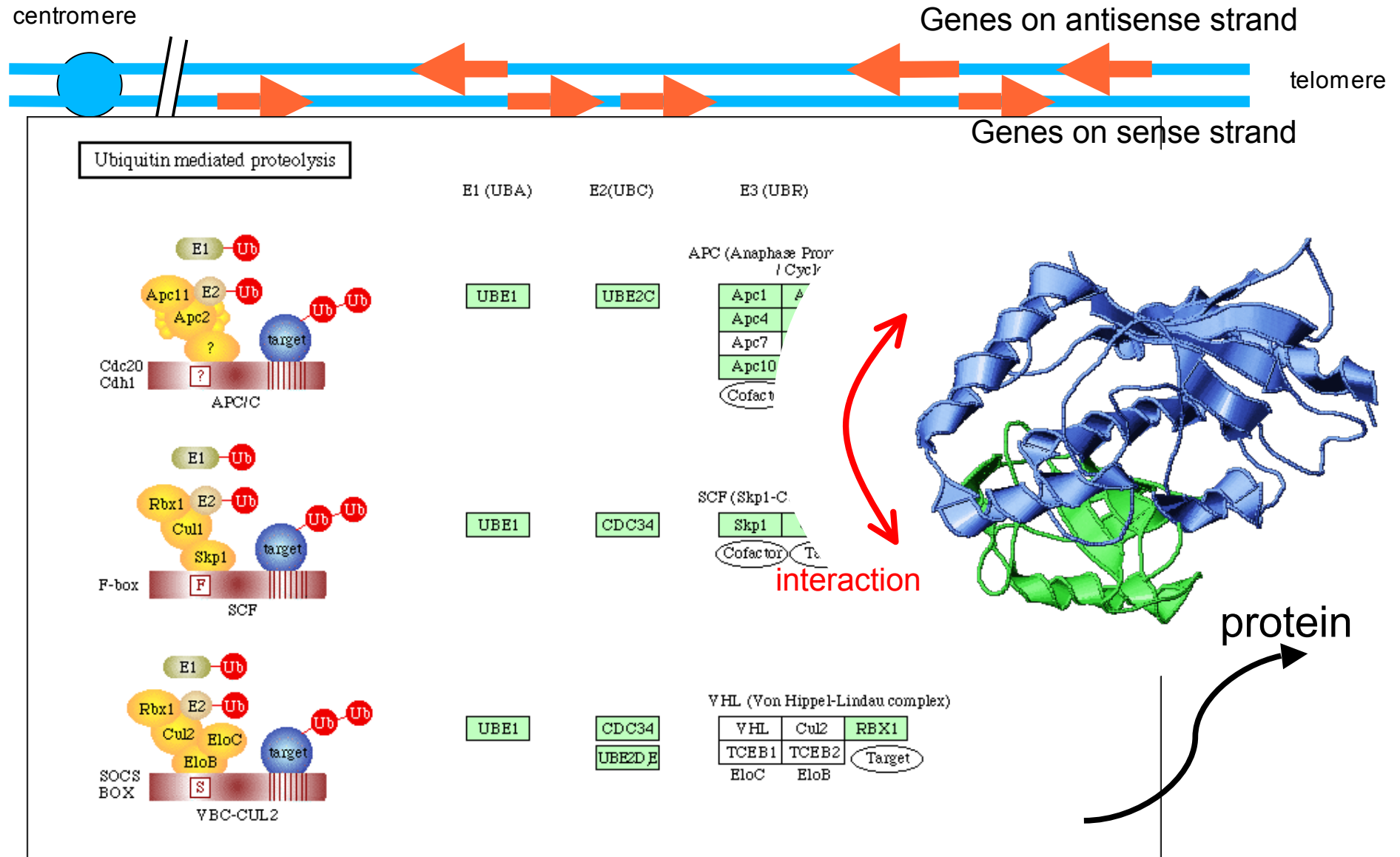
GeneMark >>>



GeneMark <<<

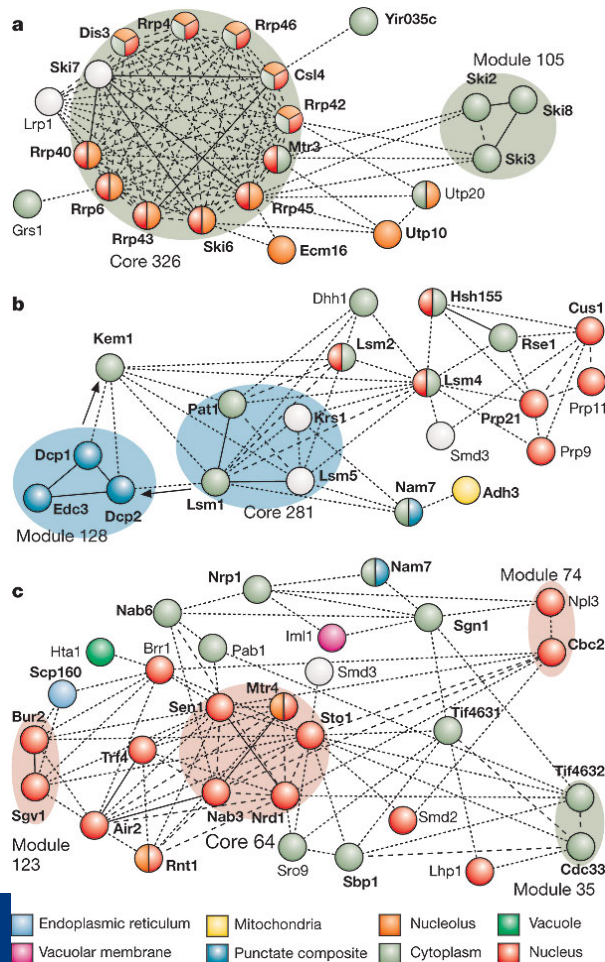


From genome to cellular function

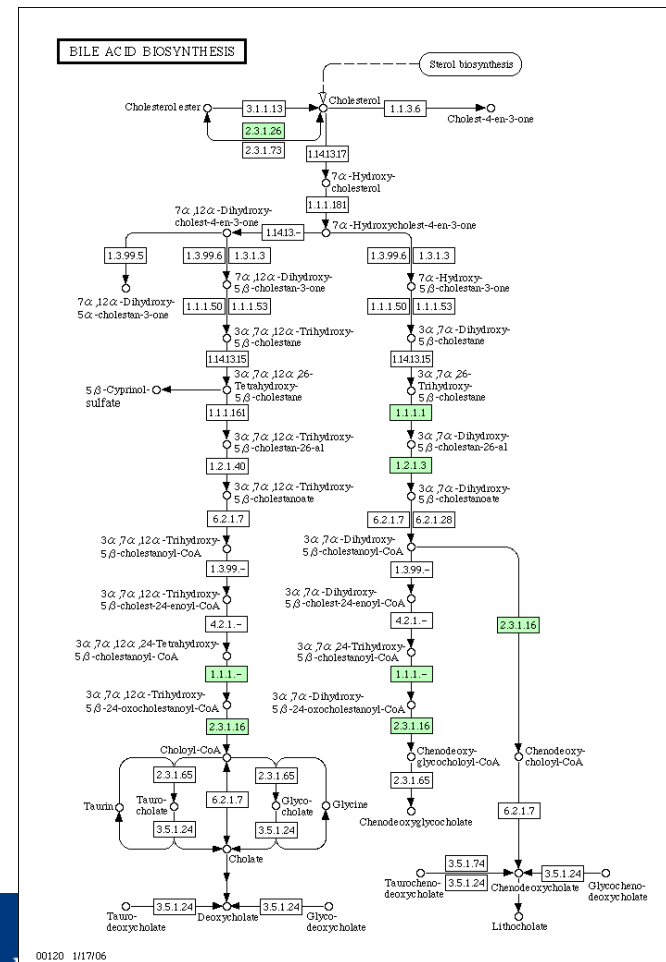


Interactions between components

interactions



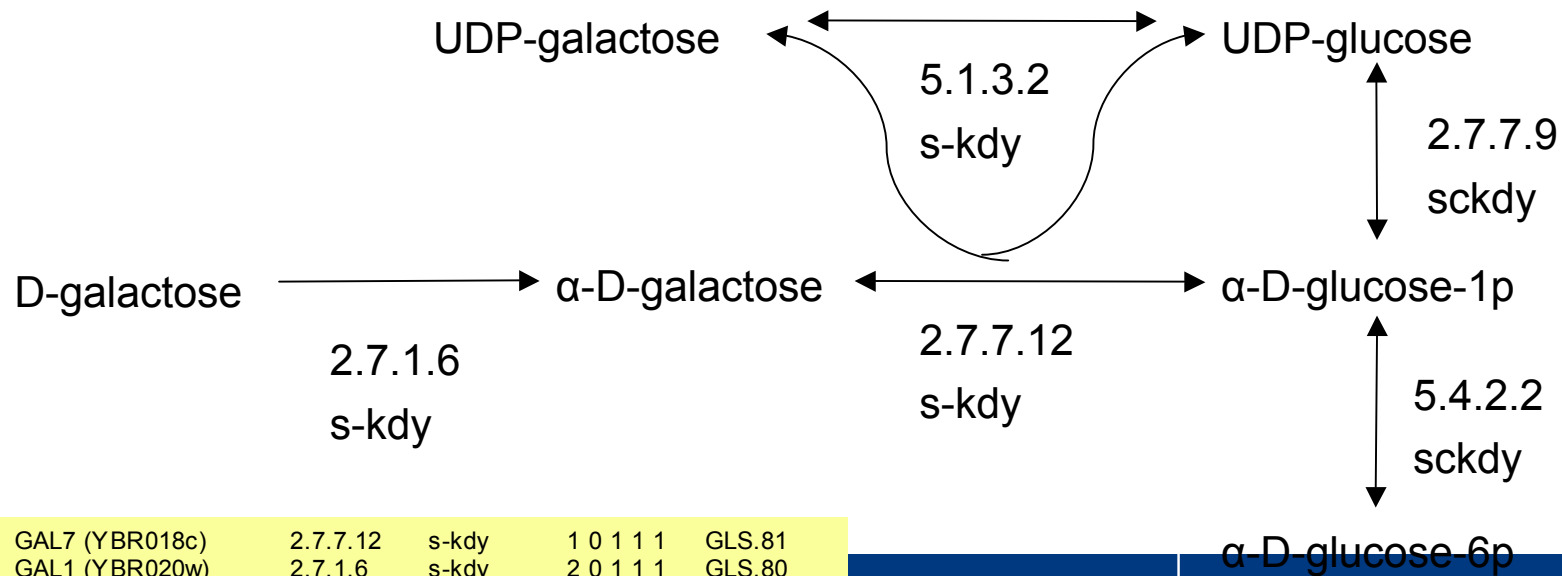
metabolism



Links from genome to function

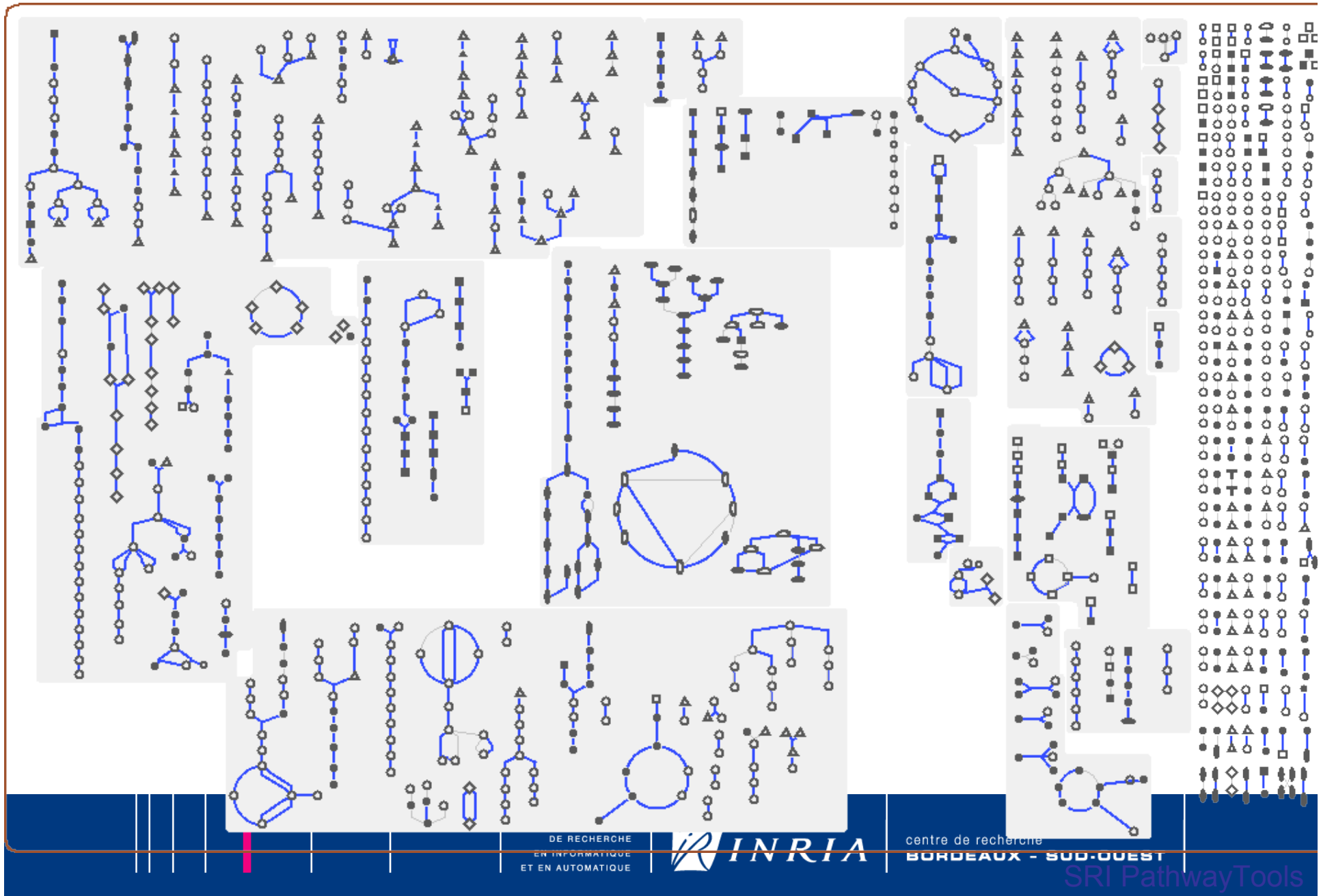
YHL012w (YHL012w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
UGP1 (YKL035w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
YHL012w (YHL012w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368
UGP1 (YKL035w)	2.7.7.9	sckdy	3 2 2 2 2	GLC.2368

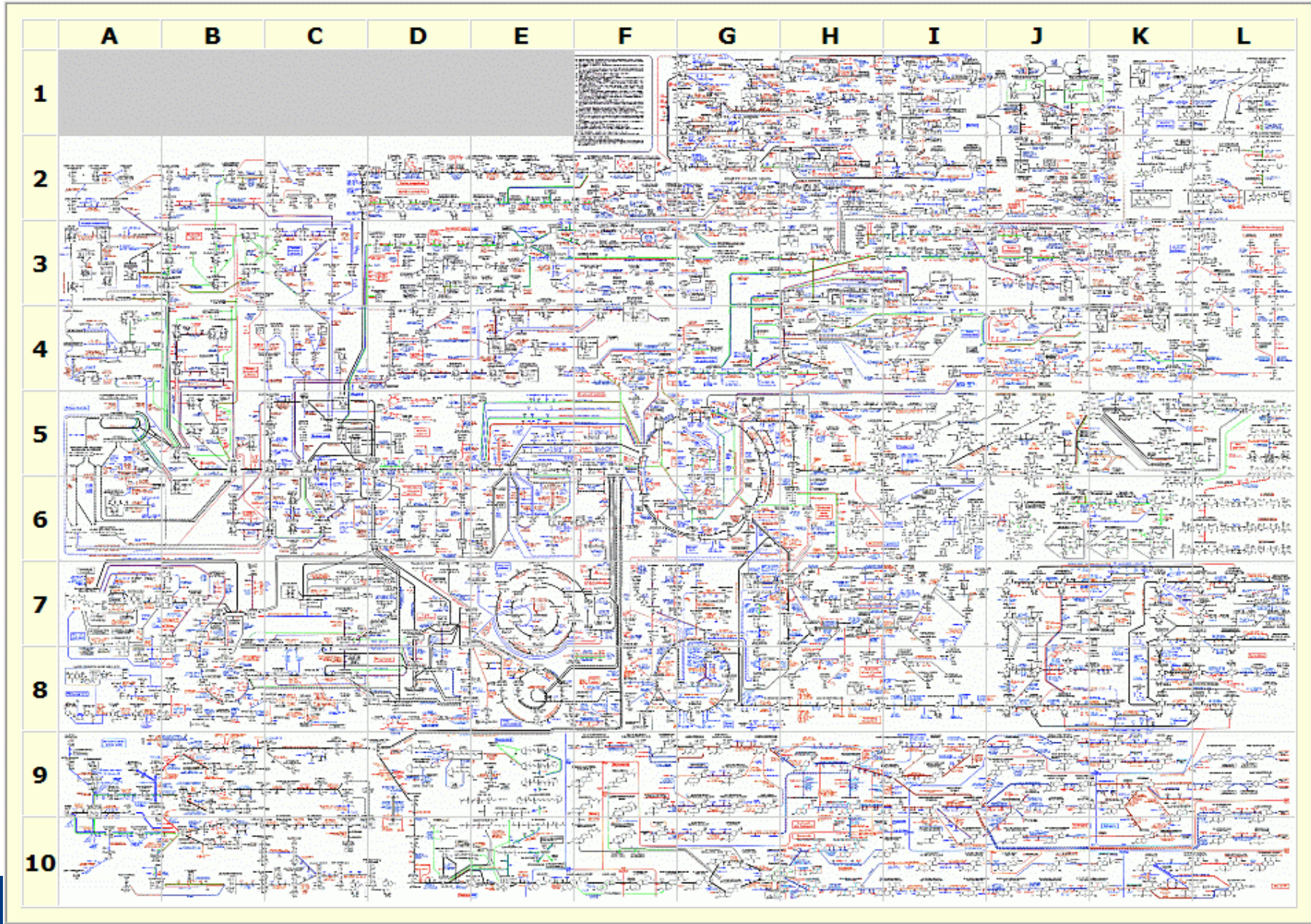
GAL10 (YBR019c) 5.1.3.2 s-kdy 3 0 1 1 1 GLS.83



GAL7 (YBR018c)	2.7.7.12	s-kdy	1 0 1 1 1	GLS.81
GAL1 (YBR020w)	2.7.1.6	s-kdy	2 0 1 1 1	GLS.80
GAL3 (YDR009w)	2.7.1.6	s-kdy	2 0 1 1 1	GLS.80

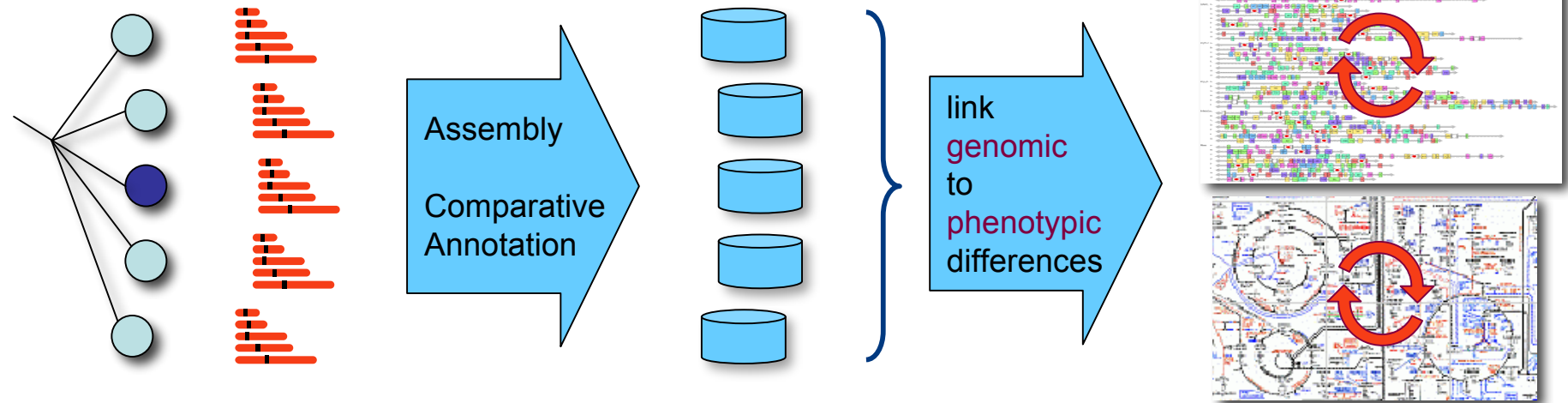
Metabolic paths in *S. cerevisiae*





MAGNOME

Algorithms and models for the genome



Related genomes compared to a reference species or strain

- cell factories
- biotechnologies
- pathogens

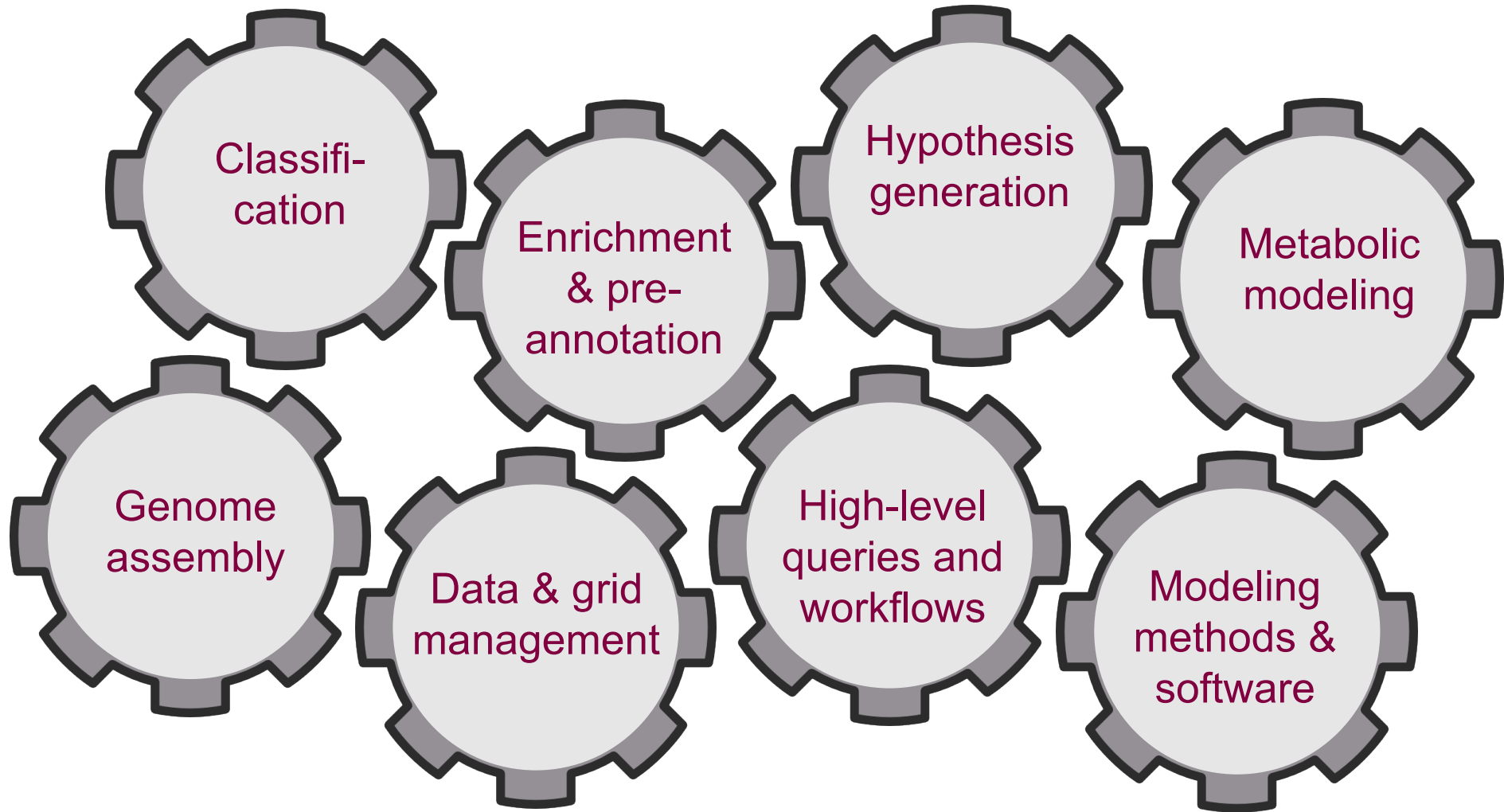
High-throughput sequencing

Genome-scale models

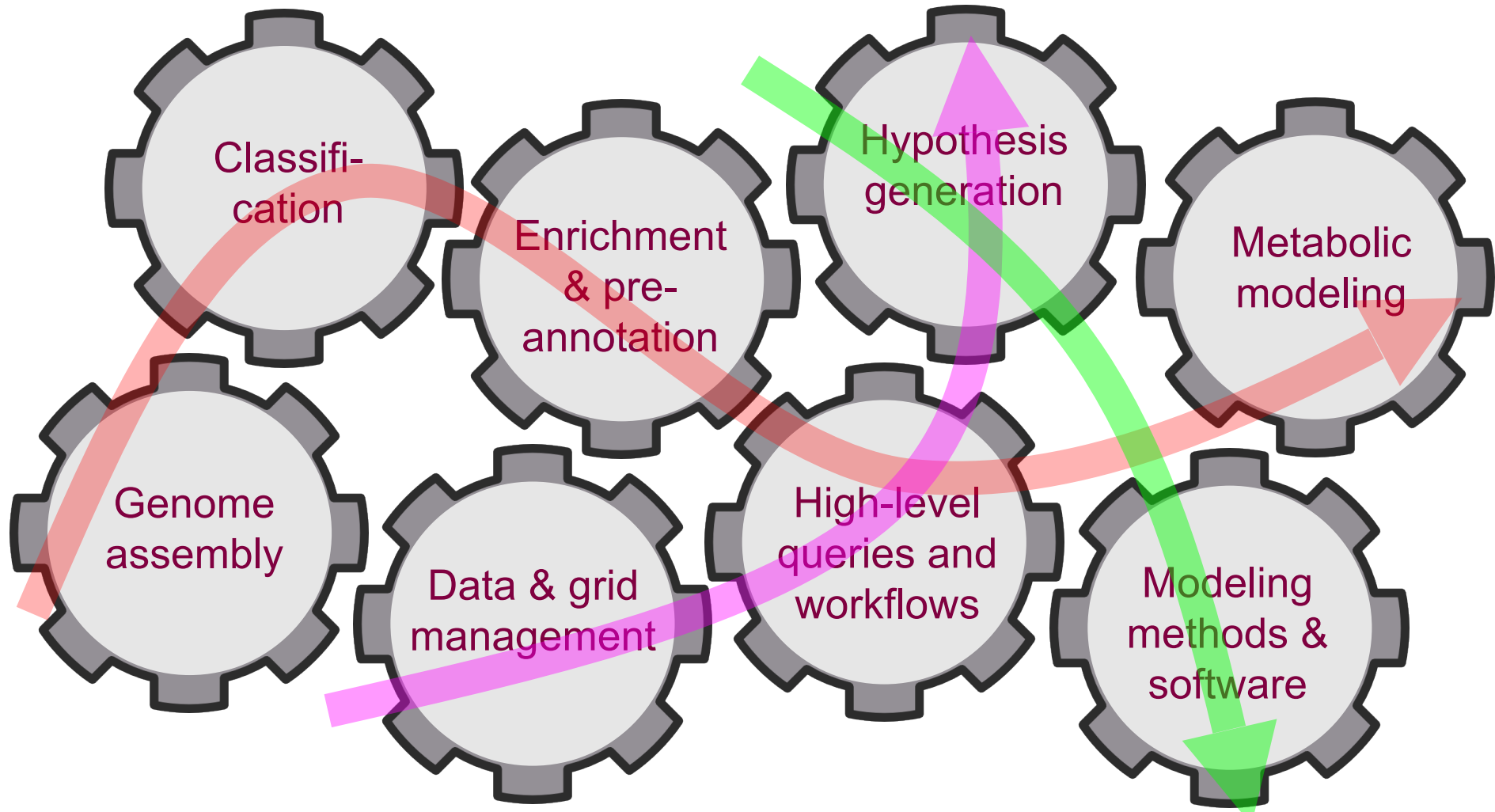
Comparative genomics
Hierarchical stochastic modeling
Classification & learning

→ <http://magnome.bordeaux.inria.fr>

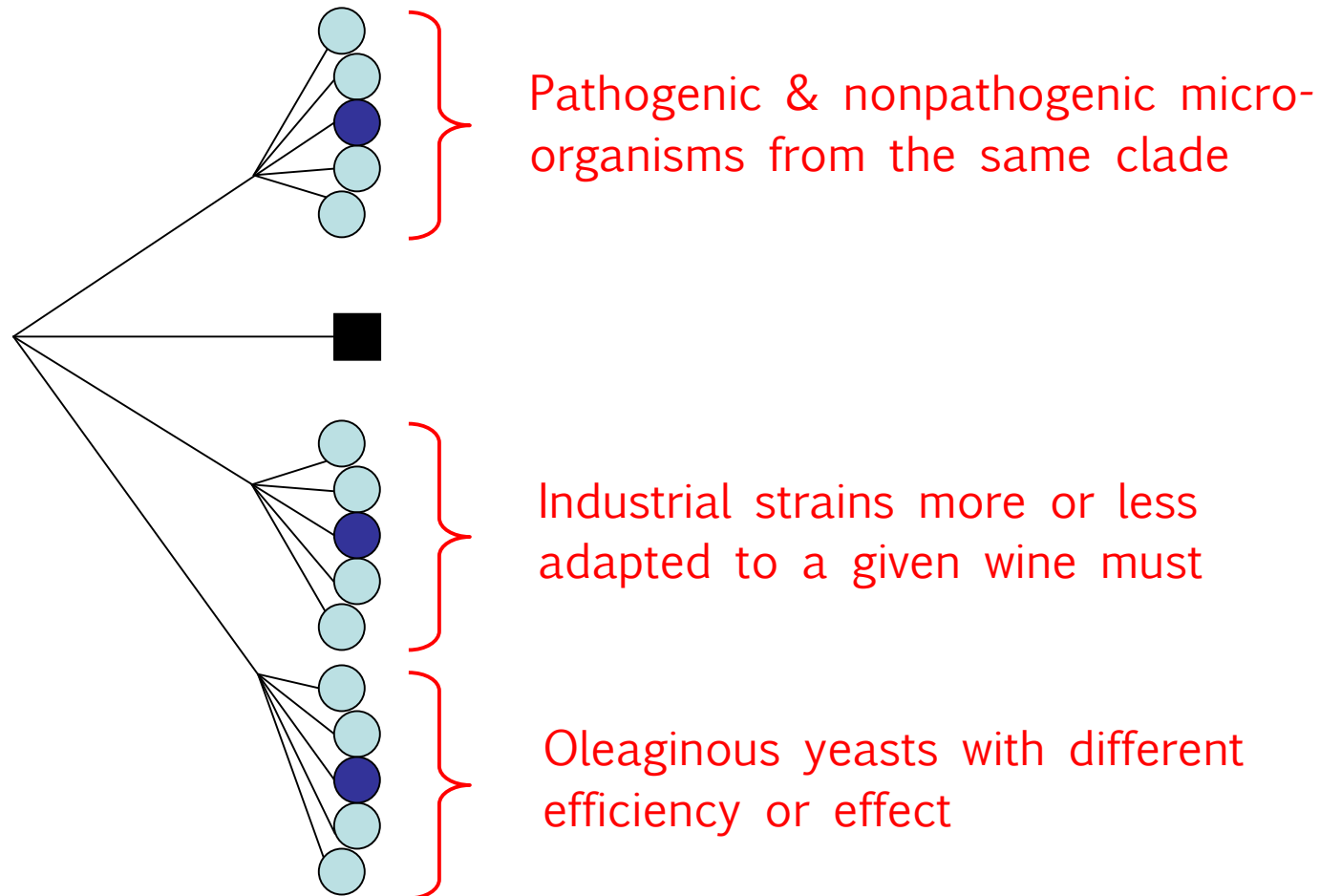
What techniques do we use?



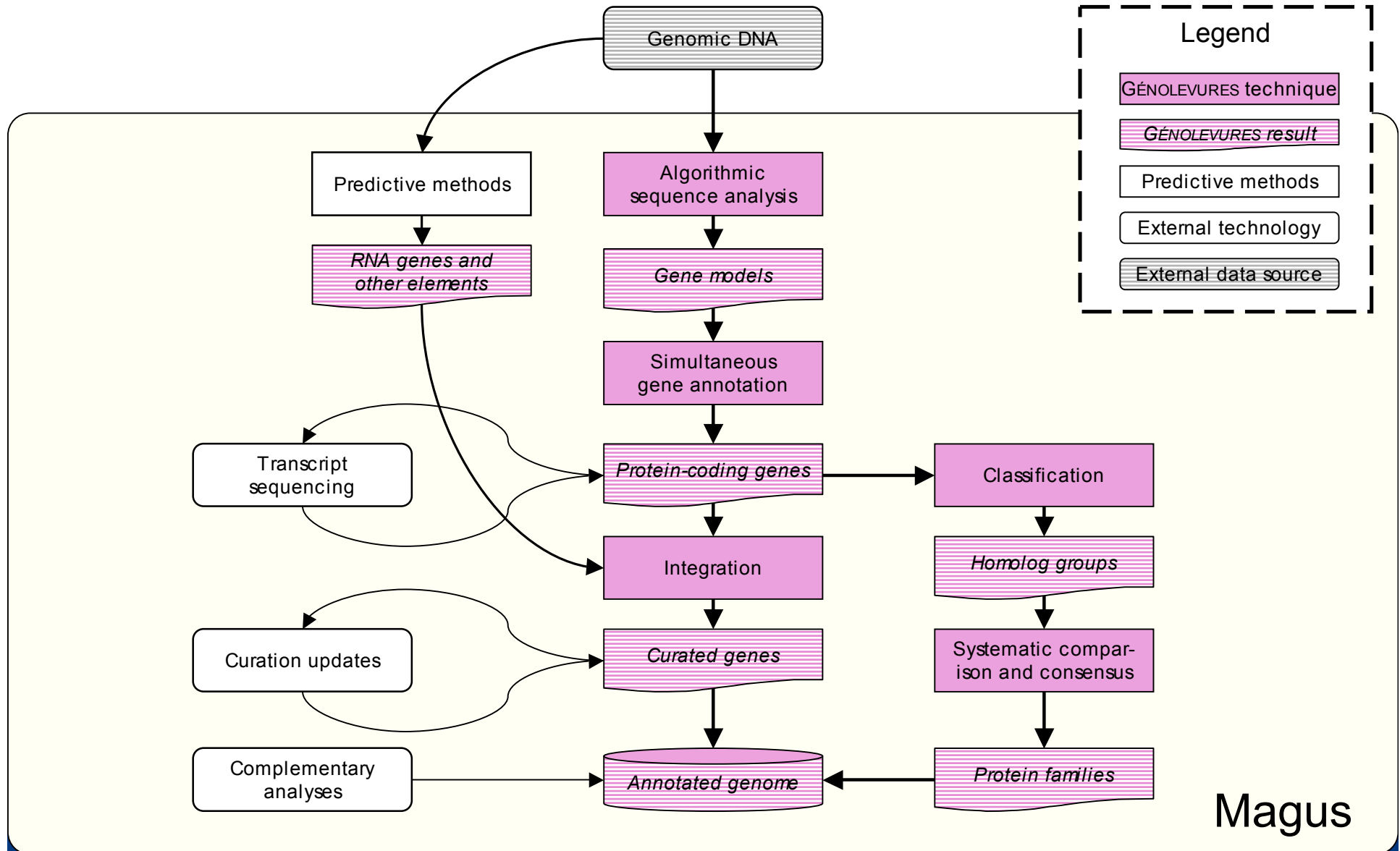
What techniques do we use?



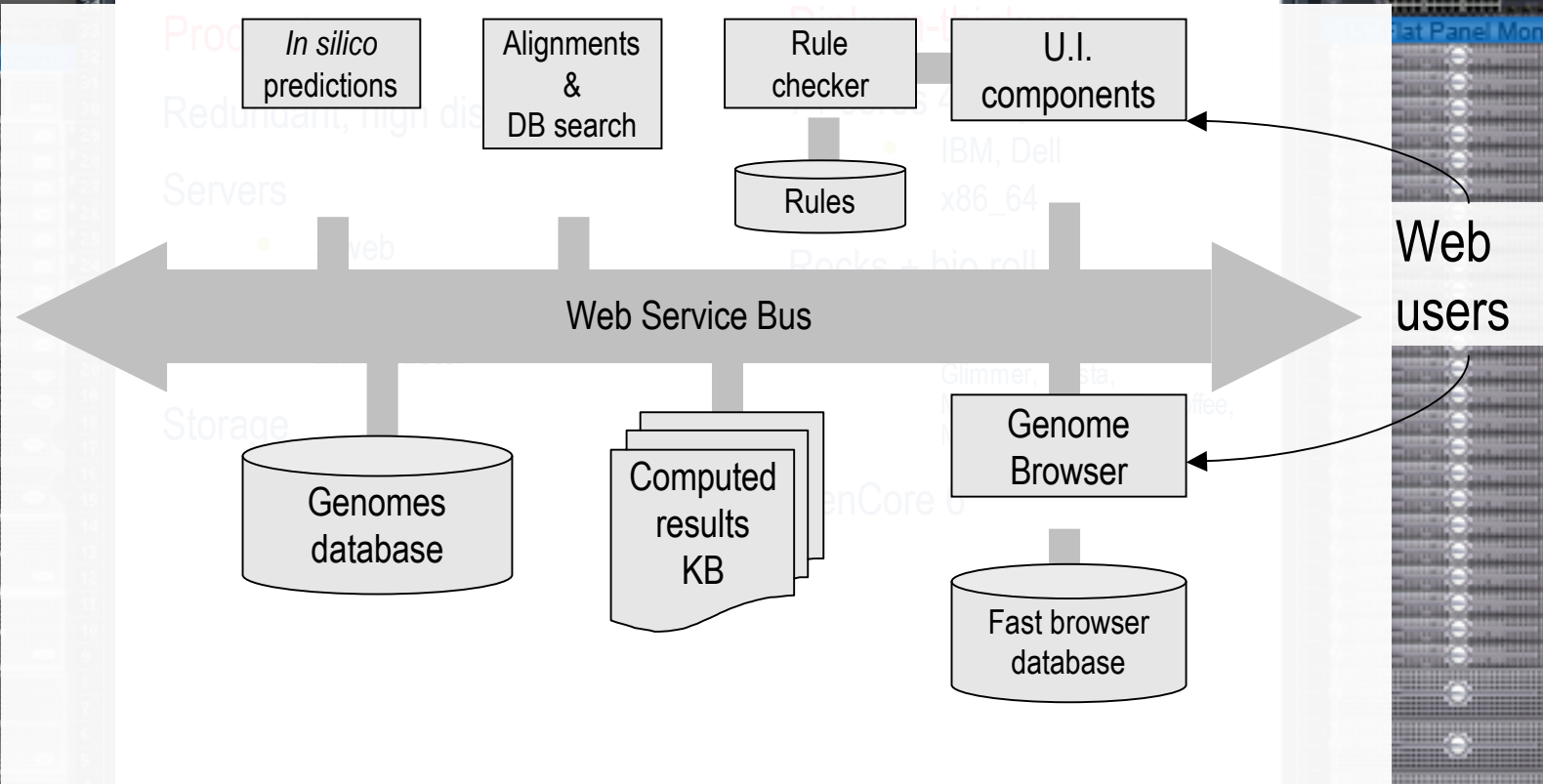
Current projects for closely related species



The Annotation Process



The "big iron"



Viewing a Locus on a Genome

Genolevures locus locus.Sakl0G.17298

http://cbl.labri.fr/Genolevures/magus/locus?id=locus.Sakl0G.17298

locus.Sakl0G.17297 ← locus.Sakl0G.17299 →

locus.Sakl0G.17298

Locus [Sakl0G](#) from 1292755 to 1294627
 Found 18 mRNA genes overlapping locus.Sakl0G.17298.

1 Validated genes

- [vSakl0G.mRNA.1321.p](#) 252 aa

0 Invalidated genes

17 Gene models

- [Sakl0G.mRNA.1321.p](#) 252 aa
- [Sakl0G.mRNA.1318.p](#) 230 aa
- [Sakl0G.mRNA.1317.p](#) 226 aa
- [Sakl0G.mRNA.1313.p](#) 221 aa
- [Sakl0G.mRNA.1314.p](#) 220 aa
- [Sakl0G.mRNA.1312.p](#) 211 aa
- [Sakl0G.mRNA.1311.p](#) 200 aa
- [Sakl0G.mRNA.1316.p](#) 194 aa
- [Sakl0G.mRNA.1309.p](#) 183 aa
- [Sakl0G.mRNA.1310.p](#) 182 aa
- [Sakl0G.mRNA.1308.p](#) 178 aa
- [SAKL-ORF3148](#) 142 aa
- [Sakl0G.mRNA.1320.p](#) 83 aa
- [Sakl0G.mRNA.1319.p](#) 74 aa
- [Sakl0G.mRNA.1315.p](#) 66 aa
- [SAKL-ORF3147](#) 55 aa
- [SAKL-ORF3149](#) 55 aa

74%

Actions

Mark locus.Sakl0G.17298 as [DONE](#)

Jump:

Sakl0G
 1291k 1292k 1293k 1294k 1295k 1296k

Locus

Genes

- vSakl0G.mRNA.1321.p
- vSAKL-ORF3150
- vSAKL-ORF3146
- vSAKL-ORF3145

Models

codon>>>

codon<<<

GeneMark

PSItblastn families

- GLC.1409
- GLS.8
- GLR.1105
- GLR.1106

ncRNA genes

BlastP Uniprot

- 63.11111111111111 YLR047C *Saccharomyces cerevisiae* 73.1481481481482 gn|GLV
- 58.8145896656535 gn|GLV|CAGL0M07942g *Candida glabrata* 52.8455284552846 tr|Q75C
- 60.7871720116618 sp|Q75CQ8 *Ashbya gossypii* 67.3228346456693 gn|GLV|KLLA0E1f
- 27.1532846715328 gn|GLV|YALI0B17292g *Yarrowia lipolytica* 55.8245083207262 t
- 69.4793536804309 gn|GLV|KLLA0E16247g *Kluyveromyces lactis* 31.383737517
- 22.6548672566372 gn|GLV|DEHA0B12122g *Debaryomyces hansenii*
- 9.38053097345133 gn|GLV|DEHA0B12122g *Debaryomyces hansenii*
- 68.5258964143426 gn|GLV|CAGL0M02

Validating a Gene Model

Genolevures gene Klth0C.mRNA.2174.m

http://cbl.labri.fr/Genolevures/magus/gene?id=Klth0C.mRNA.2174.m

KLTH-ORF14155 ← Jump: → KLTH-ORF14151

Klth0C.mRNA.2174.m

protein length is 252 aa
 Klth0C from 189741 to 190838 (antisense (-) strand)
 CDS sequence is 756 nt,
 join(complement(189741..190406),complement(190749..190838))
 wide nucleotide sequence 189541 to 191338
 GC% = , GC3% =
 Protein MW 27726.9 Da, IP 4.46, Gravy -0.218

This locus could contain a **protein-coding gene**. If this is the best predicted mRNA transcript,
 Choose this mRNA using this V_NOTE:
 highly similar to sp|P32905 Saccharomyces cerevisiae YGR214W RPS0A Protein component of the small (40S) ribosomal subunit

Quick links

Results	Homolog groups	Best-Blastp	Comments
SEQ NT	SEQ mRNA & start	SEQ AA	History

Results

Auto blastp	Auto blastp	UniProtKB blastp	UniProtKB blastp
Hemiasc blastx	GeneMark img	GeneMark lst	Interpro scan
Hemiasc tblastn	T-Coffee	TMHMM spans	

Homolog groups
 Curated homolog groups
 (no curated groups)
 Alignments with families
 PSSM for [GLS.8](#) 7e-109 unknown witness

Klth0C

189k 190k 191k

Locus

Genes
 vKLTH-ORF14155 vKlth0C.mRNA.2174.m vKLTH-OR

Models

codon>>>

codon<<<

GeneMark

PSITblastn families
 GLR.1105 GLS.8

ncRNA genes

BlastP Uniprot

64.8148148148148	gnl GLV KLLA0E16170g	Kluyveromyces lactis
52.8455284552846	tr Q75CQ0	Ashbya gossypii
58.8785046728972	gnl GLV CAGL0B02255g	Candida glabrata
61.4678899082569	YGR215W	Saccharomyces cerevisiae
46.4285714285714	gnl GLV YALI0E15312g	Yarrowia lipolytica
42.7184466019417	gnl GLV DEHA0D15840g	Debaryomyces hansenii
69.6850393700787	gnl GLV KLLA0E16214g	Kluyveromyces
75.098814229249	tr Q75CQ9	Ashbya gossypii
70.1195219123506	gnl GLV CAGL0M02849g	Candida glab
70.1612903225807	YGR214W	Saccharomyces cerevisiae
60.3703703703704	gnl GLV YALI0A18205g	Yarrowia lip
64.367816091954	gnl GLV DEHA0B14702g	Debaryomyces

Annotating Homolog Groups

Genolevures group vGLR.641

http://cbi.labri.fr/Genolevures/magus/group?id=vGLR.641

PSITblastn families

K11a0B

199k 198k 197k 196k 195k 194k 193k 192k 191k 190k 189k 188k 187k 186k 185k 184k 183k 182k 181k

Locus

Genes

vKLLA-ORF9717 vKLLA-ORF9721 vKLLA-ORF9724 vKLLA-ORF9730 vKLLA-ORF9734

vKLLA-ORF9718 vKLLA-ORF9722 vKLLA-ORF9727

browse

Validate these annotations // Select all Clear all Reset // Copy group define Copy GO terms *

Homolog group annotation vGLR.641 (change name here)

Define (for the family as a whole)

homolog group vGLR.641 derived from GLR.641

GO terms (add terms here)

kinase activity catalytic activity

Gene 1 – YPL214C in groups vGLR.641, vC.7284_1, vC.7284, GLR.641.
p|P41835 Saccharomyces cerevisiae YPL214c THI6 thiamin-phosphate pyrophosphorylase, member vGLR.641

kinase activity catalytic activity

Gene 2 – vCAGL0E05808g in groups vGLR.641, vC.7284_1, vC.7284, GLR.641, GLR.641.
highly similar to sp|P41835 Saccharomyces cerevisiae YPL214c THI6 thiamin-phosphate pyrophosphorylase start by similarity, member vGLR.641

kinase activity catalytic activity

Gene 3 – vKLLA-ORF9724 in groups vGLR.641, vC.7284_1, vC.7284, GLR.641.
similar to sp|P41835 Saccharomyces cerevisiae YPL214c THI6 bifunctional enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-

kinase activity catalytic activity

Gene 4 – vKLTH-ORF1764 in groups vGLR.641, vC.7284_1, vC.7284.
enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-hydroxyethylthiazole kinase activities, member vGLR.641

kinase activity catalytic activity

Gene 5 – vSAKL-ORF15337 in groups vGLR.641, vC.7284_1, vC.7284.
enzyme with thiamine-phosphate pyrophosphorylase and 4-methyl-5-beta-hydroxyethylthiazole kinase activities, member vGLR.641

kinase activity catalytic activity

Gene 6 – vYALI0C15554g in groups vGLR.641, vC.7284, GLR.641, GLR.641.
phosphate pyrophosphorylase and hydroxyethylthiazole kinase start by similarity, member vGLR.641

kinase activity catalytic activity

Validate these annotations // Select all Clear all Reset // Copy group define Copy GO terms *

2190k 2180k

Locus

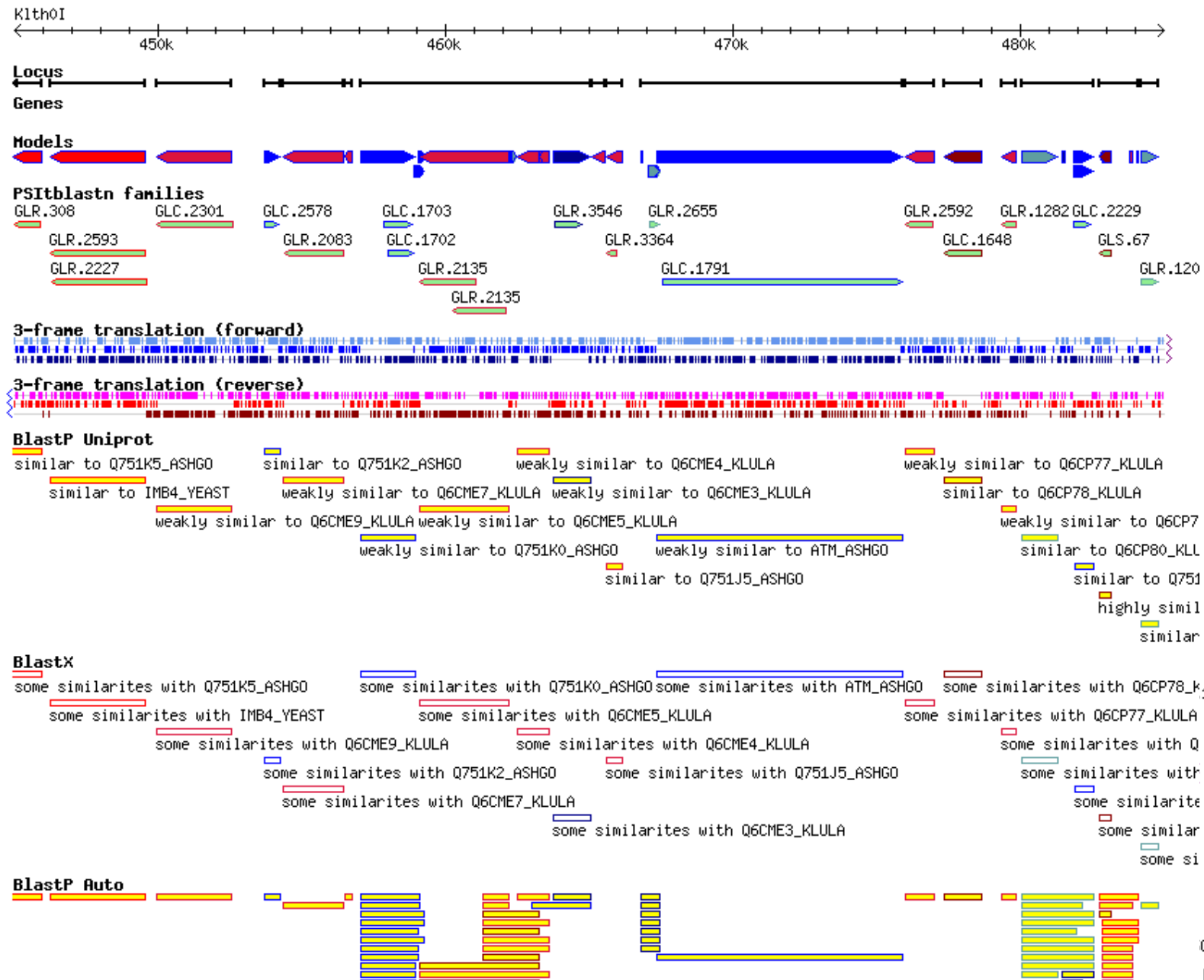
Genes

vYALI0C15554g

PSITblastn families

browse

Browsing a genome region



Which one are you?

The hard part

Data

A solved problem

Push button



The hard part

Biological

—

otherwise
not interesting

Results

Algorithmic

—

otherwise
not interesting



MAGNOME

Algorithms and models for the genome

Permanent

- D. Sherman DR INRIA HDR
- P. Durrens* CR CNRS HDR
- E. Bon* MCF U.Bordeaux 2
- T. Martin* IR CNRS

**biologist or pluridisc. training*

Non permanent

- A. Garcia, IJD INRIA
- A. Goulielmakis, IE ANR
- N. Loira, PhD CONICYT (4/4)
- R. Assar, PhD INRIA (3/3)
- A. Sarkar, PhD EMMA (3/3)
- N. Golenetskaya, PhD INRIA (2/3)
- R. Issa, PhD PAB Syria (1/4)
- Alexander Makarov (intern DRI)
- Anna Zhukova (intern DRI)



Some open questions

Collaborative genome annotation

- Interaction design and human-computer interfaces

Mathematical modeling

- Hierarchical stochastic models of fermentation
- Metaheuristic optimization fo evolutionary scenarios

High-performance computing

- Deploying grid-aware web services
- Distributed NoSQL databases

Contact

- David Sherman david.sherman@inria.fr
- <http://magnome.bordeaux.inria.fr>



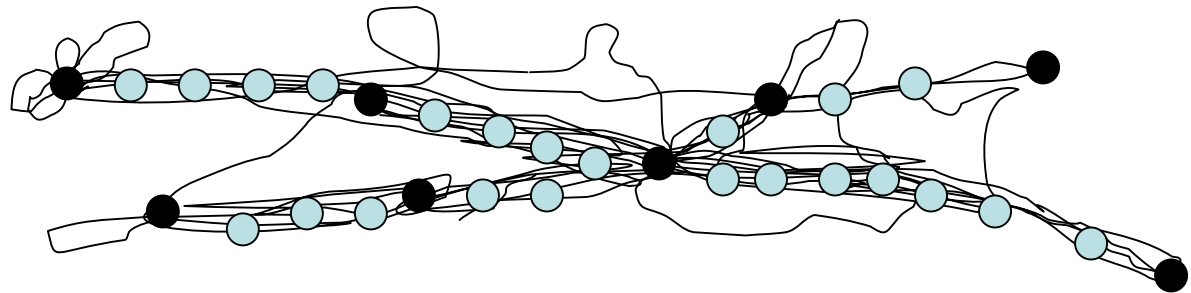
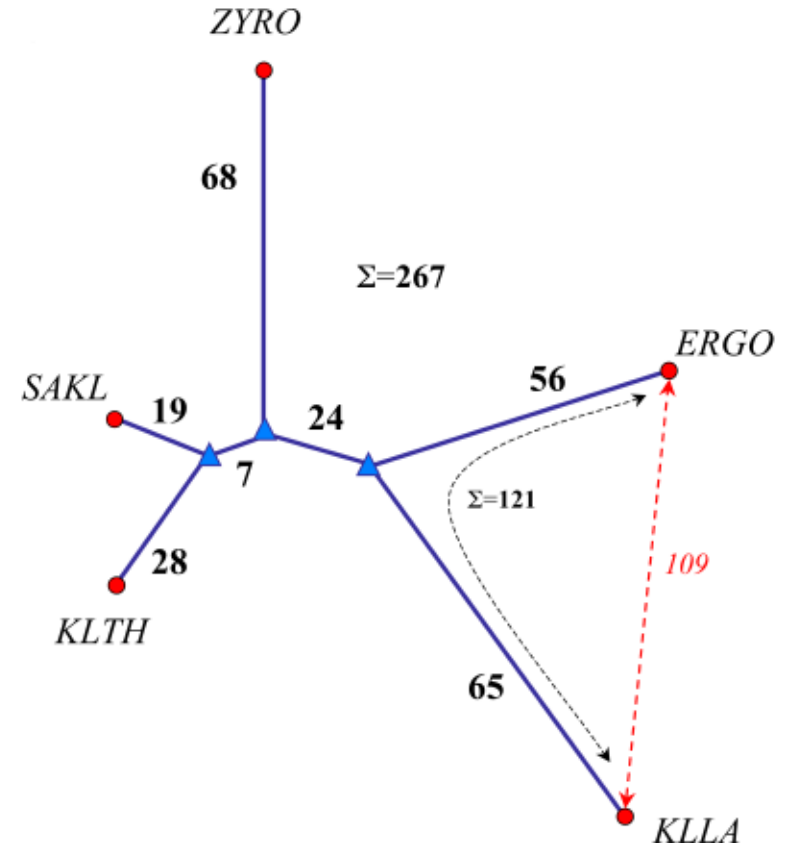
Two examples

Finding median genomes

- Population
- Stochastic local search
- Experimental results

Finding parsimonious scenarios

- Ant colony optimization
- Hybridization with random walk
- Experimental results



Finding median genomes

[Goëffon *et al* GECCO 2008]

A *chromosome* $\pi = (\pi_1, \dots, \pi_m)$ is represented by a sequence of signed gene markers whose sign indicates their relative direction on the chromosome. A size- n *multichromosomal genome* Π is defined as a set of chromosomes $\{\pi^1, \dots, \pi^N\}$ such that $\sum_i |\pi^i| = n$. Markers take their value from the set of ordinals $1, \dots, n$; no given marker appears in more than one chromosome.

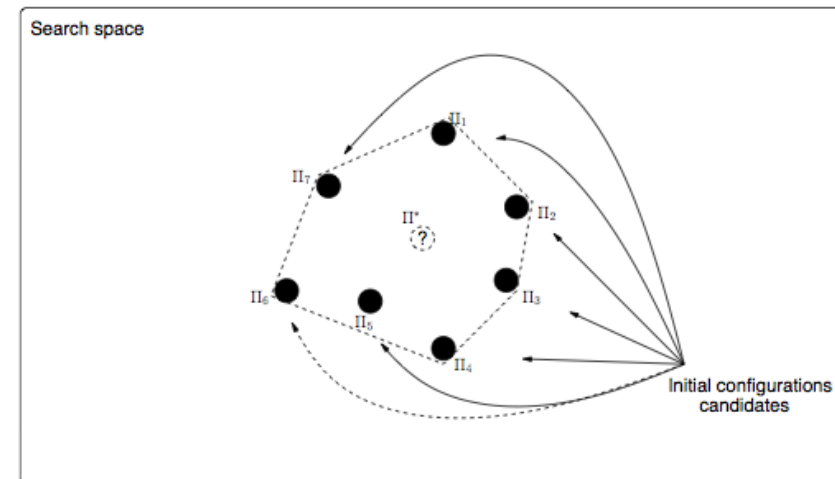
Given a set of size- n genomes $\{\Pi_i\}$ and a genome distance function d , an instance of the combinatorial minimization problem *MGP* is defined by two elements $\langle \tau_n, \phi \rangle$:

1. a search space, τ_n , composed of the set of all possible size- n genomes, and
2. an objective function $\phi : \tau_n \rightarrow \mathbb{N}$ (*score*) defined by $\phi(\Pi) = \sum_{\Pi_i} d(\Pi, \Pi_i)$.

A *median genome* for a given set of genomes $\{\Pi_i\}$ is a genome Π that minimizes $\phi(\Pi)$. Every optimal solution to *MGP* is a median genome.

Population-based local search

No crossing-over



1. the *initial configuration* is taken from $\{\Pi_i\}$,
2. the *evaluation function* is the same as the objective function ϕ : the rearrangement distance d ,
3. the *neighborhood relation* we call \mathcal{R}^1 is a 1-step rearrangement: $\mathcal{R}^1(\Pi) = \{\Pi' \in \tau_n, d(\Pi, \Pi') = 1\}$,
4. the *move strategy* is a *first-improve selection* (FI) which accepts better and equivalent configurations (*side-walk mechanism*, SW [16]), given a specified number of iterations *nbit*.

Probabilistic neighbor selection

Need to select the most pertinent neighbors at each step

- Simultaneous descent
- Dependant replications

Probabilistic neighborhood

- Encourages adjacencies which are not, or are less represented in the population to be broken
- Replaces as a function of the proportional representation of the broken adjacencies

Algorithm 1

Require: $\{\Pi_1, \dots, \Pi_k\}$: a set of k multichromosomal genomes of size n ; l, k : the size of the population; $nbit$: the number of descent iterations.

Ensure: an approximate median genome set $\hat{\mathcal{P}}$

let \mathcal{P} be the multiset of the current genomes population, which initially contains l copies of each Π_i .

let $numit \leftarrow 0$ be the number of performed local search iterations

while $numit < nbit$ **do**

for all $\Pi \in \mathcal{P}$ **do**

loop

 randomly select $\Pi' \in \mathcal{R}^1(\Pi)$

break with probability $p(\Pi, \Pi', \mathcal{P} \setminus \{\Pi\})$

end loop

if $\phi(\Pi') \leq \phi(\Pi)$ **then**

$\Pi \leftarrow \Pi'$

end if

end for

$numit \leftarrow numit + 1$

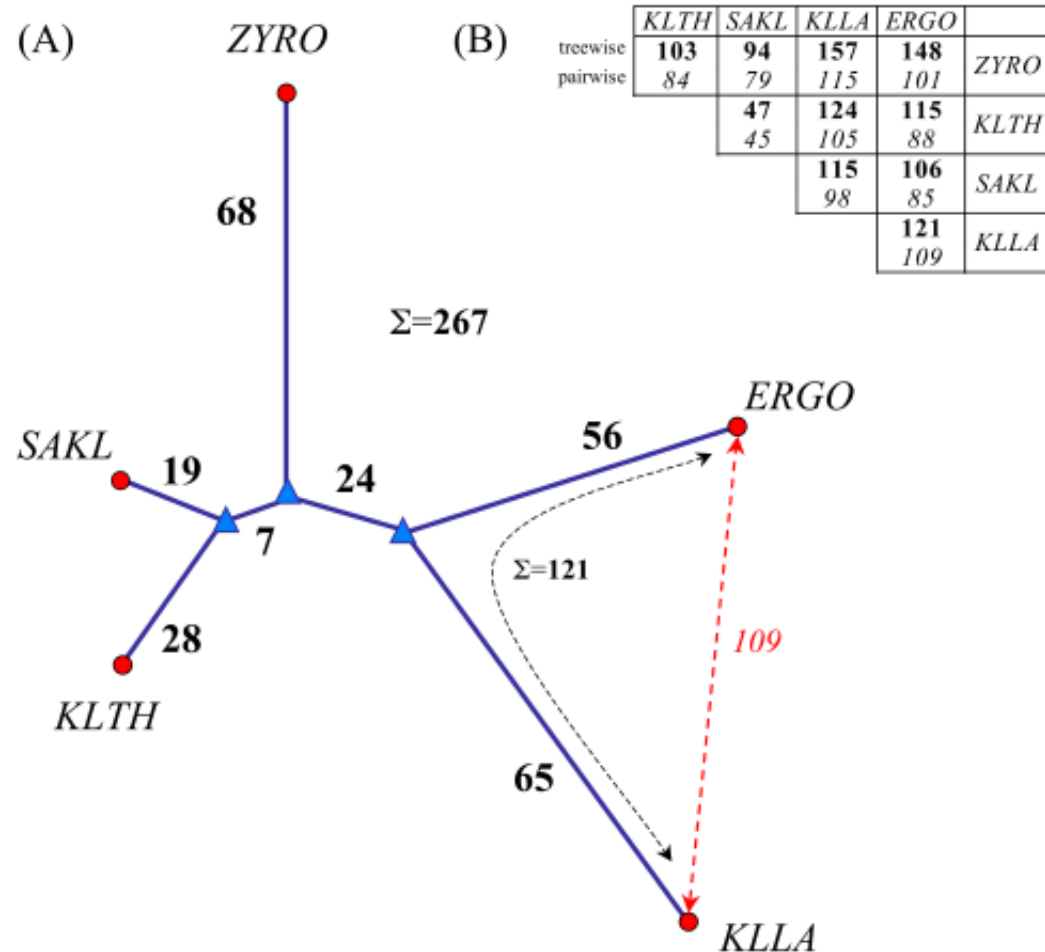
end while

return $\hat{\mathcal{P}} = \operatorname{argmin}_{\Pi \in \mathcal{P}} \phi(\Pi)$

$$p(\Pi, \Pi', \mathcal{P} \setminus \{\Pi\}) = 1 - \frac{|\{\Pi'' \in \mathcal{P} \setminus \{\Pi\}, (\mathcal{A}(\Pi) \setminus \mathcal{A}(\Pi')) \cap \mathcal{A}(\Pi'') \neq \emptyset\}|}{|\mathcal{P}| - 1}$$



Protoploid Hemiascomycete yeasts



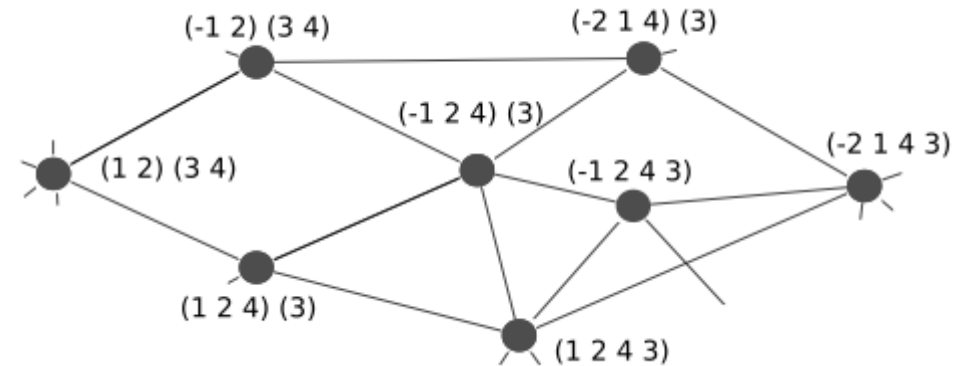
[Souciet *et al* Genome Res. 2009]

Multiple scenarios

[Vyahhi *et al* GECCO 2009]

Explore solution space

- No **one** scenario is best
- Suboptimal solutions may be more biologically plausible



ACO formulation

- $\Delta\tau_\varepsilon = 1/(l - d + 1)$
- Evaporation rate ρ
- Edge detection threshold ε
- Select k random neighbors (except in the endgame)

DEFINITION 1. A **scenario** of length m between two size- n multichromosomal genomes Π and Γ is a sequence $S = (\Pi_0, \dots, \Pi_m)$ s.t.

- $\forall i \in [0, m], \Pi_i$ is a size- n multichromosomal genome,
- $\Pi_0 = \Pi$ and $\Pi_m = \Gamma$, and
- $\forall i \in [0, m - 1], d(\Pi_i, \Pi_{i+1}) = 1$.

If $m = d(\Pi, \Gamma)$, then S is a **parsimonious scenario**. A sequence S s.t. $m = d(\Pi, \Gamma) + \alpha$ without cycles, that is, where $\Pi_i \neq \Pi_j \forall i, j \in \{0, \dots, m\} (i \neq j)$, is called an α -**parsimonious scenario**

