Les Apicomplexes: un challenge pour la Bioinformatique et la biologie théorique

Développements théoriques et méthodes numériques pour les analyses comparatives de génomes et protéomes biaisés

Olivier Bastien

INRA Institut National de la Recherche Agronomique
cea
Financé par ANR
GENE·IT

Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

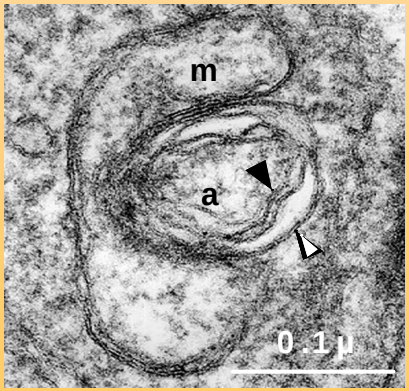
- Conclusion générale

Plasmodium falciparum:
l ’agent du paludisme

- Le paludisme: 2,5 millions de morts par an

- Agent infectieux: *Plasmodium falciparum*
(→ séquençage complet depuis Octobre 2002)

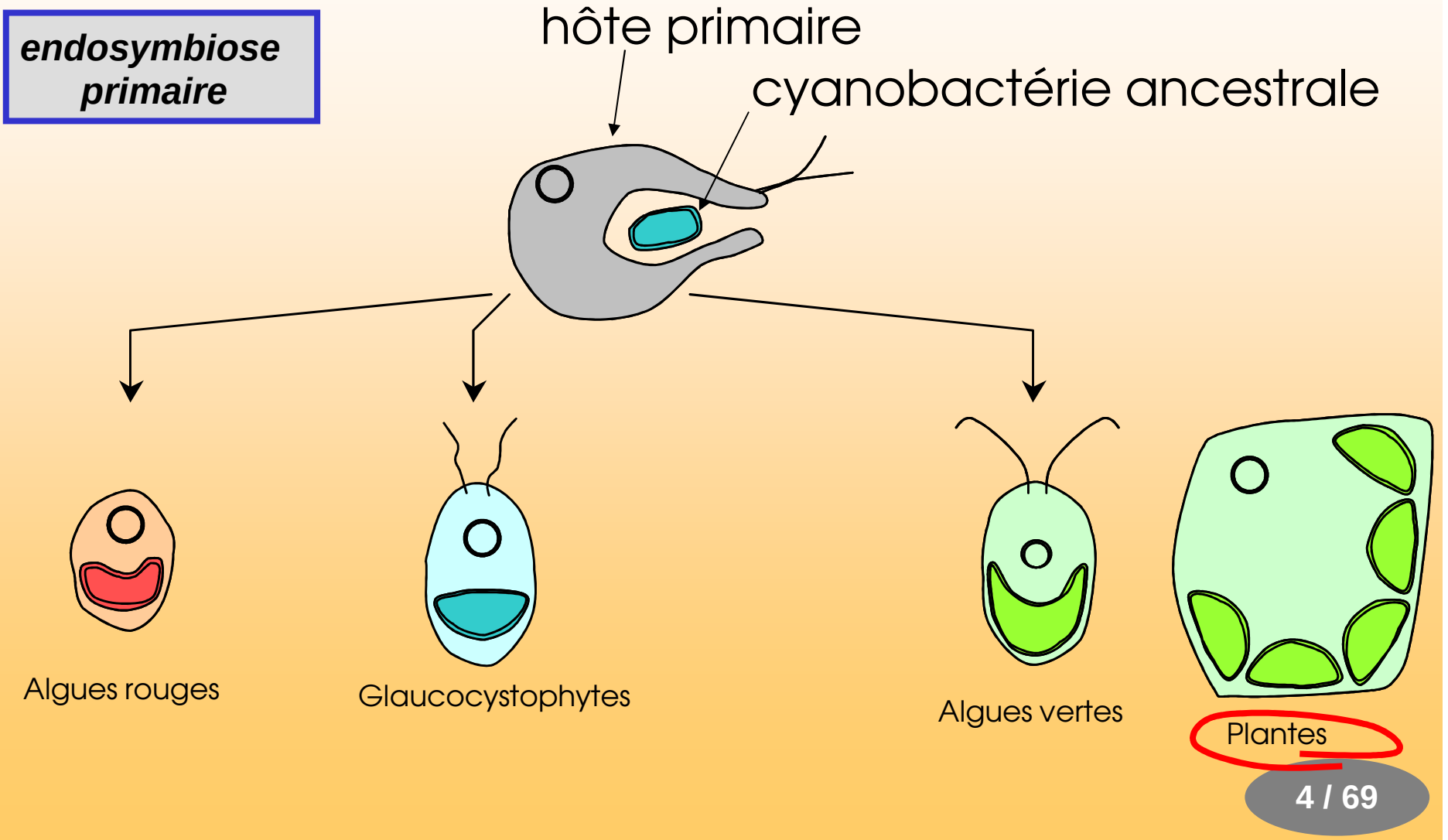
5300 ORF prédits
3000 gènes de fonctions inconnues
2000 sans homologues



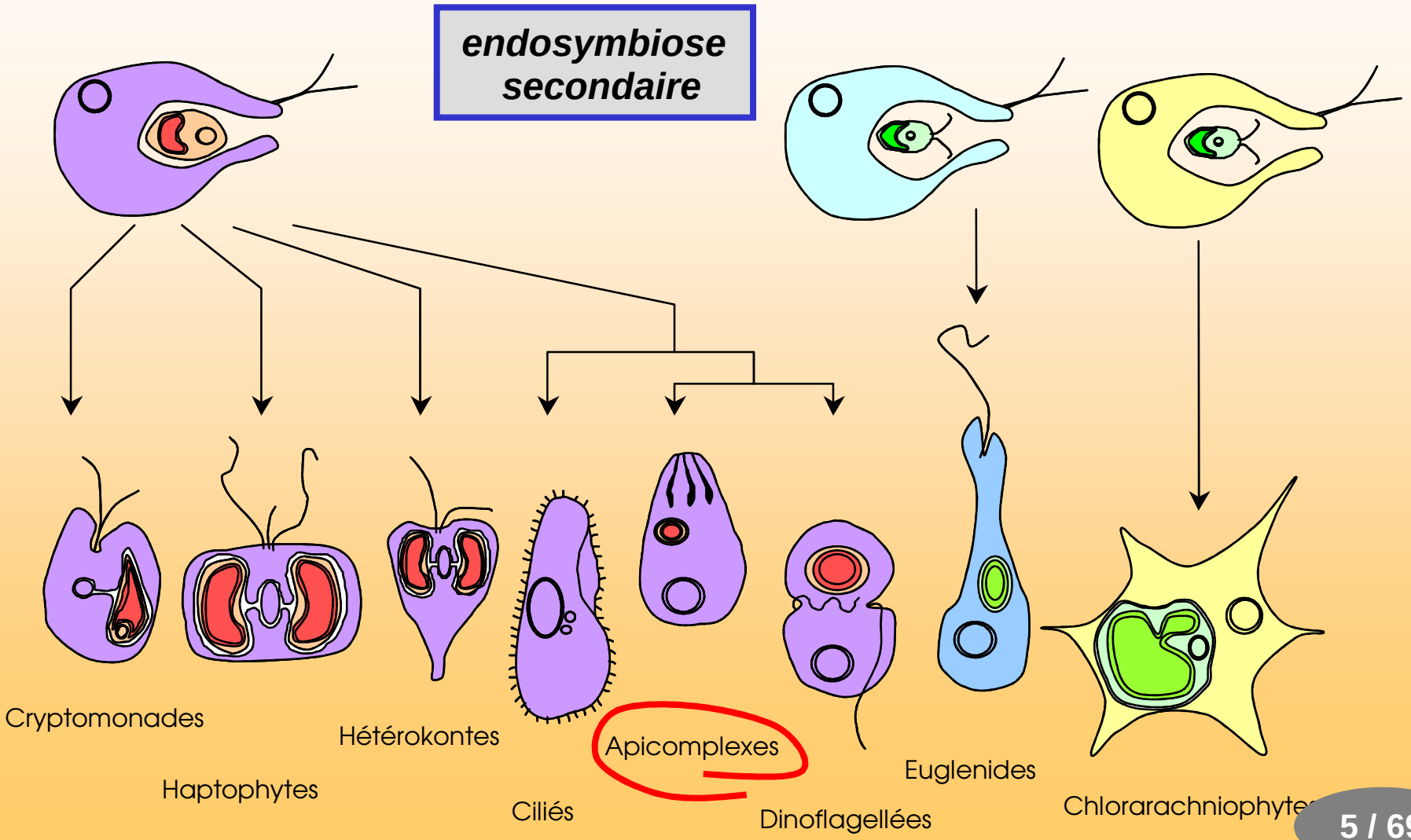
-*Plasmodium falciparum* contient des structures typiquement végétales

Plasmodium falciparum:
Une histoire évolutive complexe (1)

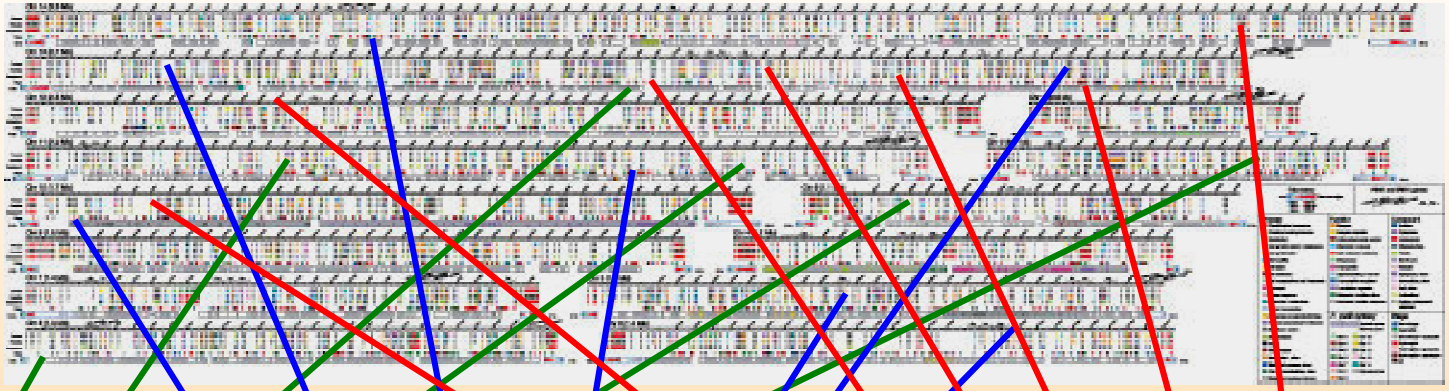
endosymbiose primaire



Plasmodium falciparum:
Une histoire évolutive complexe (2)



Plasmodium falciparum: *un génome nucléaire composite*



issus de la cyanobactérie ancestrale

issus du premier hôte

issus de l'hôte final

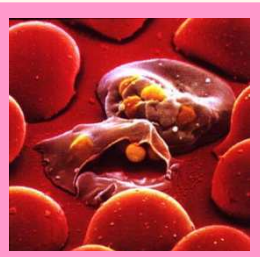
issus de l'algue rouge

=> Sous-génome végétal ?

Médicaments herbicides

Analyse comparée des génomes d'*Arabidopsis thaliana*, de *Plasmodium falciparum* et de l'homme

Proche des plantes

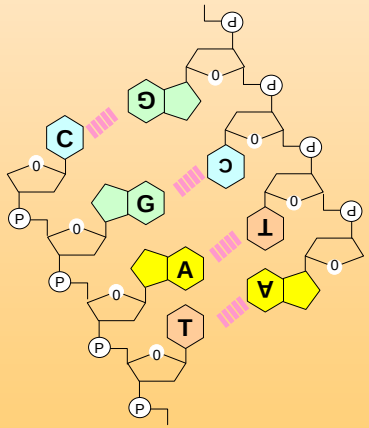


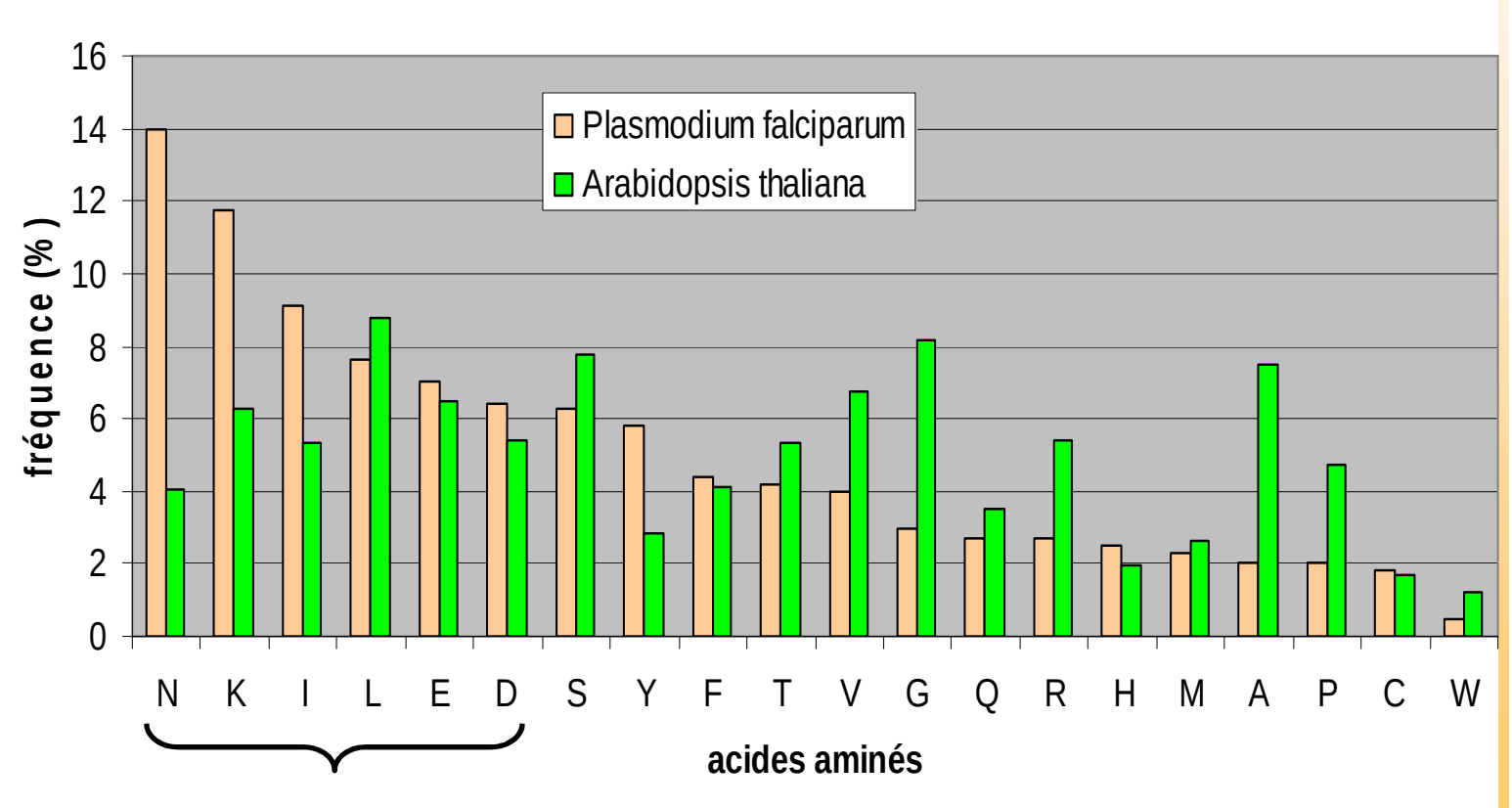
séquences de Plasmodium 82% A+T

Proche des mammifères

Le génome de *Plasmodium falciparum* est biaisé en A-T

- *Plasmodium falciparum* : **82% d'adénosine-thymidine**
- *Arabidopsis thaliana* : 50% d'adénosine-thymidine





Objectifs

Le biais en aminoacides entraîne-t-il:

1- une incertitude sur l'estimation de la significativité d'un score d'alignement?

2- une incertitude sur la qualité des alignements?

3- Le biais en aminoacides est-il une conséquence du biais en acides nucléiques?

4- Quel cadre théorique et pratique pour la recherche du sous-génome végétal de *Plasmodium*?

Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

- Conclusion générale

La parenté évolutive de séquences primaires est mesurée grâce à des alignements (1)

Postulat fondamental de l'analyse de séquences:

1- Les séquences de deux molécules de fonctions apparentées vont en général présenter des ressemblances

2- Réciproquement, deux molécules dont les séquences présentent des ressemblances ont probablement des fonctions apparentées

La parenté évolutive de séquences primaires est mesurée grâce à des alignements (2)

Blocks de taille ≥5

Block1 *Block2* *Block3* *Block4* *Block5*

```
DRT1_ARATH: 242 MVEDIISNGNVKNDRTGTGTLSKFGCQMKFNLRRSFPLLTTKRVFWRGVVEELLWFISGS 301
                ++ DI+ NGN ++DRTG G LSKFG  MKF+L + FPLLTTK++F RG++EELLWFI
DRTS_PLAFK: 331 IIYDIMMNGNKQSDRTGVGVLSKFGYIMKFDLSQYFPLLTTKKLFLRGIIEELLWFIRGE 390
```

	Block1	***Block2***	***Block3***	***Block4***	***Block5***
taille des blocks	6	5	5	10	10
identité (%)	67	100	80	80	80

Principe de la mesure d'un alignement (1)

- On attribue à chaque alignement un score

- Pour tenir compte de la proximité de certains acides aminés (en terme de propriétés physico-chimiques ou autres), on utilise un **matrice de similarité** S de dimension 20×20 qui tient compte de toutes les combinaisons possibles de paires d'acides aminés

- S_{jk} ,ou $S(j,k)$, est la qualité de l'alignement de l'acide aminé j avec l'acide aminé k

Principe de la mesure d'un alignement (2)

i dans la séquence **requête**
aligné avec ***j*** dans la séquence **sujette**,
avec une fréquence q_{ij}

$$S_{ij} = \lambda . \log \left(\frac{q_{ij}}{p_i p_j} \right)$$

Principe de la mesure d'un alignement (3)

On a alors:

$$\begin{cases} q_{ij} \geq p_i\, p_j \Rightarrow S_{ij} \geq 0 & \text{Substitution favorable} \\ q_{ij} \leq p_i\, p_j \Rightarrow S_{ij} \leq 0 & \text{Substitution défavorable} \end{cases}$$

Principe de la mesure d'un alignement (4)

Le score global de l'alignement de deux séquences de longueur L est alors calculé par:

$$score = \sum_{k=1}^{L} S(a_k, b_k)$$

global

pour chaque résidu

L'alignement optimal est celui qui maximise le score

Évaluation de la pertinence d'un score (1): Le modèle de Karlin & Altschul (1990)

1- Classiquement: estimation de la probabilité d'obtenir un score avec le modèle de Karlin & Altschul (1990):

$$P(X \geq s) = 1 - \exp(-K.m.n.e^{-\lambda s})$$

2- Les hypothèses du modèle:

- Les distributions des aminoacides dans les deux séquences comparées "ne sont pas trop dissimilaires "

- Les séquences ont des tailles "comparables"

=> Hypothèses violées quand on compare des séquences biaisés (Plasmodium falciparum)

Évaluation de la pertinence d'un score (2): La *Z-value* de Lipman-Pearson (1985)

Technique permettant d'évaluer la robustesse d'un score $s(a,b)$ entre deux séquences a et b

1- Génération de 1000 permutations aléatoires de b => b^*

2- Pour chaque permutation, alignement de a avec b^* => $s(a,b^*)$

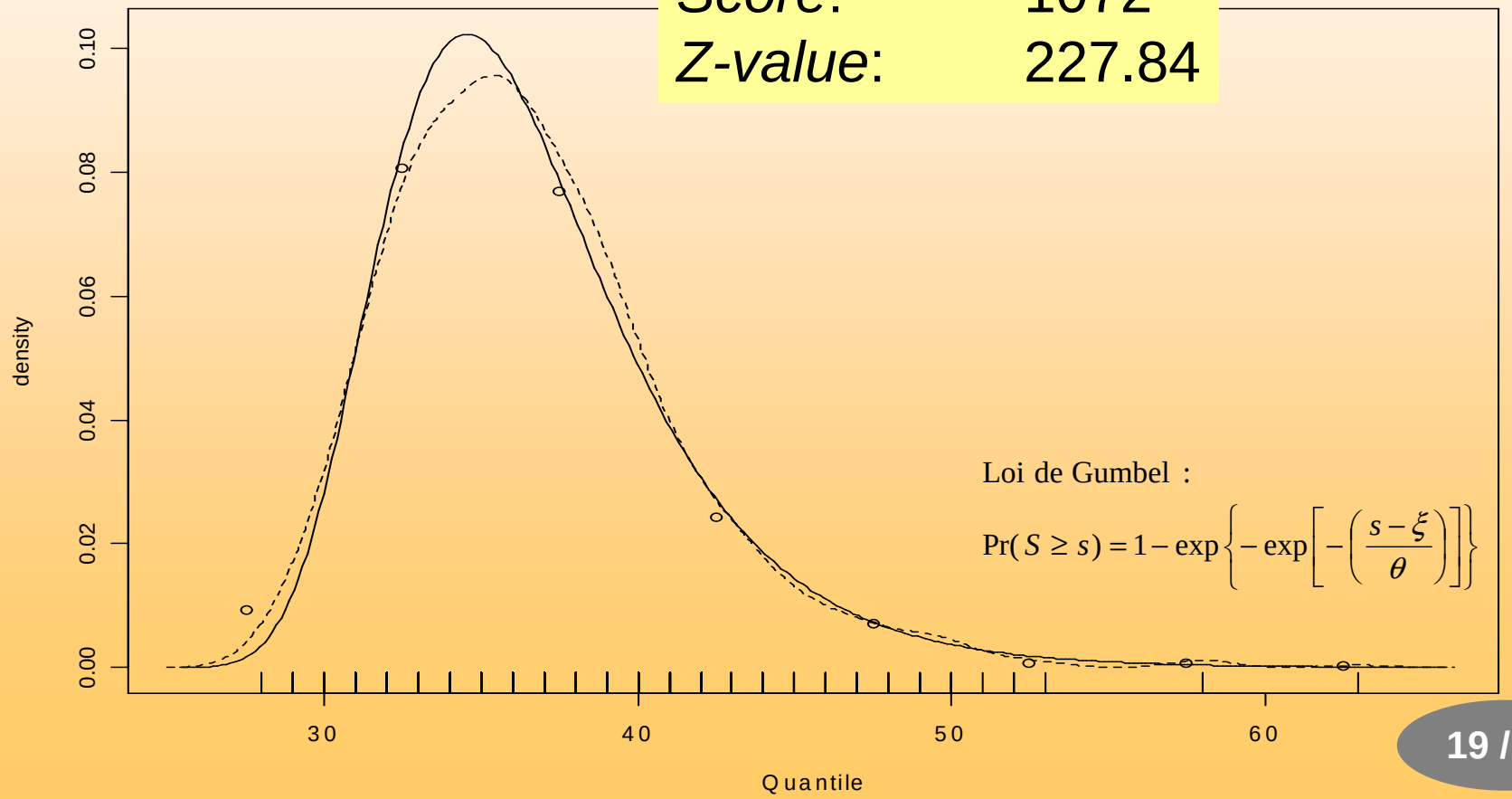
3- On observe la distribution des 1000 $s(a,b^*)$. Où se situe $s(a,b)$ dans cette distribution?

$$Z\text{–}value = \frac{s(a,b) - E[S(a,b^*)]}{\sigma}$$

Evaluation de la pertinence d'un score (3):

Exemple: alignement smith-waterman de la DHFR
d'*Arabidopsis thaliana* et de *Plasmodium falciparum*

Score: 1072
Z-value: 227.84



Loi de Gumbel :

$$\Pr(S \geq s) = 1 - \exp\left\{-\exp\left[-\left(\frac{s-\xi}{\theta}\right)\right]\right\}$$

Evaluation de la pertinence d'un score (4): Pertinence de la *Z-value*

1) La Z-value est une mesure utilisée pour évaluer un résultat d'alignement (fourni par la méthode de Smith & Waterman ou celle de Blast).

2) Il n'existait pas de démonstration pour justifier théoriquement l'utilisation de ce paramètre.

3) Les différentes expériences ont montrées que les alignements dont la *Z-value* est supérieure à 8 sont des alignements statistiquement peu probables et que très souvent, l'homologie entre les séquences est avérée.

4) l'étude du cas extrême de l'analyse comparée des protéomes de P. falciparum et A. thaliana nécessite une démonstration sur la pertinence de ce paramètre.

Signification théorique du Z-Score (1)
théorème TULIP

On se donne deux séquences réelles $a=(a_1a_2\ldots a_m)$ et $b=(b_1b_2\ldots b_n)$ pour lesquelles on a $s=s(a,b)$, le score d'alignement entre a et b tel que défini par Altschul et al. (1990) et par Smith et Waterman (1981). Soit b^* une séquence aléatoire correspondant à la séquence b randomisée et $P(S(a,b^*)\geq s(a,b))$ la probabilité qu'une séquence b^* aléatoire ait un score avec a supérieur ou égal à $s(a,b)$.

Théorème: *Quelque soit la distribution de la variable aléatoire $S(a,b^*)$, on a la relation:*

$$s\geq E[S(a,b^*)]+k\sigma \Rightarrow P(S(a,b^*)\geq s)\leq \frac{1}{k^2}$$

Corrolaire 2 de TULIP

De façon général:

$$P(S(a,b^*) \geq s(a,b)) \leq \frac{1}{z(a,b^*)^2}$$

$$\frac{1}{z(a,b^*)^2} = T\text{–}value$$

Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

- Conclusion générale

Rappel sur les notions de proximité

Similarité

On appelle similarité dans E toute fonction $f(x,y): E \times E \rightarrow \Re^+$ telle que :

i) $\forall x \in E, \forall y \in E, f(x,x) = \max_y \left(f(x,y) \right)$

ii) $\forall x \in E, \forall y \in E, f(x,y) = f(y,x)$

Dissimilarité

On appelle dissimilarité dans E toute fonction $f(x,y): E \times E \rightarrow \Re^+$ telle que :

i) $\forall x \in E, \forall y \in E, f(x,y) = 0 \Leftrightarrow x = y$

ii) $\forall x \in E, \forall y \in E, f(x,y) = f(y,x)$

Distance

On appelle distance dans E toute fonction $d(x,y): E \times E \rightarrow \Re^+$ telle que

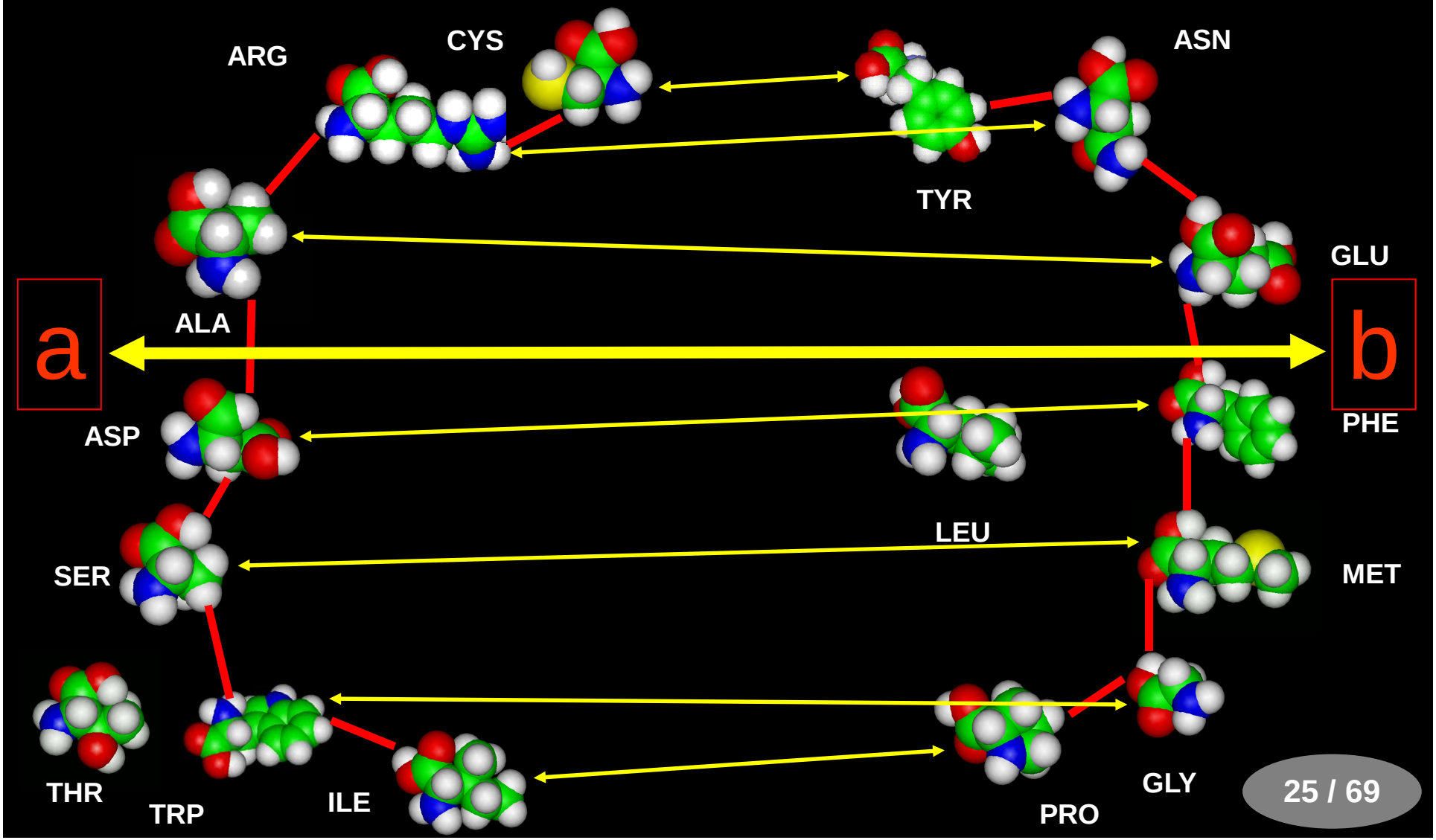
i) $\forall x \in E, \forall y \in E, d(x,y) = 0 \Leftrightarrow x = y$

ii) $\forall x \in E, \forall y \in E, d(x,y) = d(y,x)$

iii) $\forall x \in E, \forall y \in E, \forall z \in E, d(x,z) \leq d(x,y) + d(y,z)$

De l'espace des acides aminés à l'espace des séquences

quelle fonction de proximité adopter?



Les matrices de substitutions

-L’espace des amino acides est mal connu: beaucoup de facteurs complexes

- Longueur et taille de la chaîne latérale
- poids moléculaire
- solubilité dans l’eau
- pK
- Nature du groupement chimique radical

- Nécessité de mesurer une proximité entre acide aminé dans cet espace

- Solution empirique formulée par Dayhoff et al. (1978) et Henikoff and Henikoff (1992)

$$s(i,j)=\log \frac{q_{ij}}{\pi_i \pi_j}$$

La théorie de l'information: les bases (1)

Comment transmettre des données à moindre coût (bonne compression) mais avec un bon niveau de fiabilité (redondance)?

<u>Bell laboratories</u>

Hartley, 1928
Shannon, 1948

La théorie de l'information: les bases (2)

- La réception d'un message n'est susceptible d'apporter de l'information que si son contenu n'est pas connu à l'avance du destinataire

---

- L'information apportée par un événement est donc liée à la surprise que sa réalisation procure

PROBLEME : surprise est difficilement chiffrable

L'idée de Shannon (1948) : lier l'information apportée par un événement E à sa probabilité de réalisation

La théorie de l'information: les bases (3)

- **Incertitude (au sens de Hartley (1928))** liée à un événement E:

$$h(E) = -\log(P(E))$$

, mesure l'information sur le système apportée par l'occurrence de E

- Si E et F sont indépendant, on a $h(E \cap F) = h(E) + h(F)$

- ***Information mutuelle entre évènements***: information apportée par l'occurrence d'un événement F sur la possible occurrence de E

$$I_{F \to E} = h(E) - h(E / F)$$

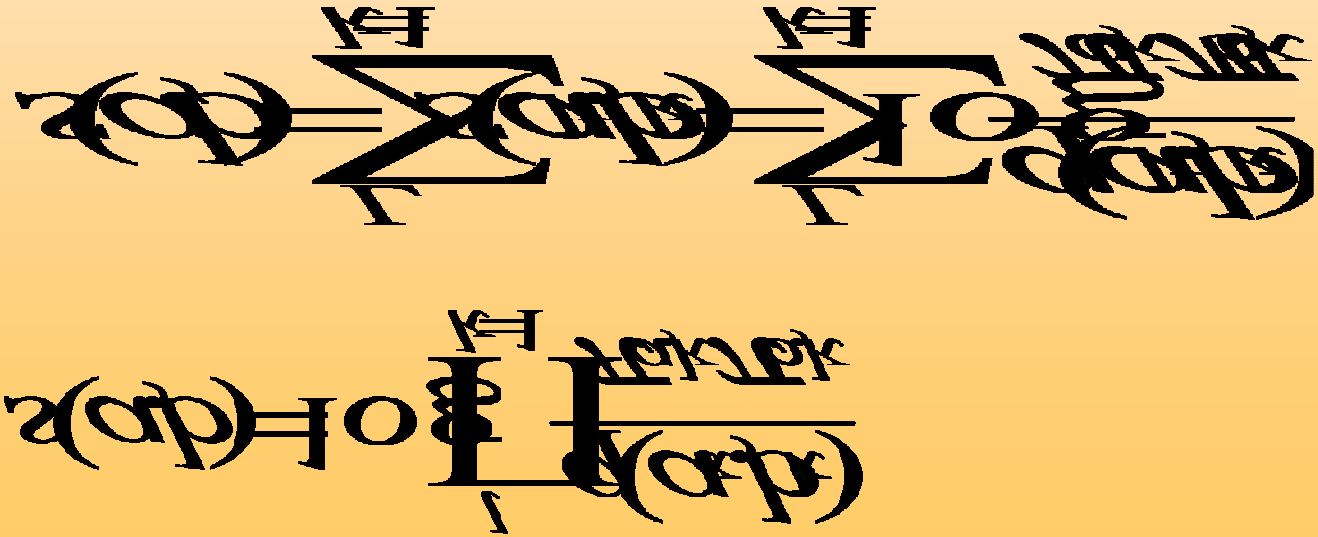
- On montre que $I_{F \to E} = I_{E \to F} = I(E;F)$

information mutuelle entre E et F

Les matrices de substitution sont des matrices d'*informations mutuelles* (3)

$$s(\dot{a},\dot{b}) = I((\dot{a};\dot{b}))$$

La façon dont on effectue la mesure « information mutuelle » sous-tend l'hypothèse d'indépendance des résidus

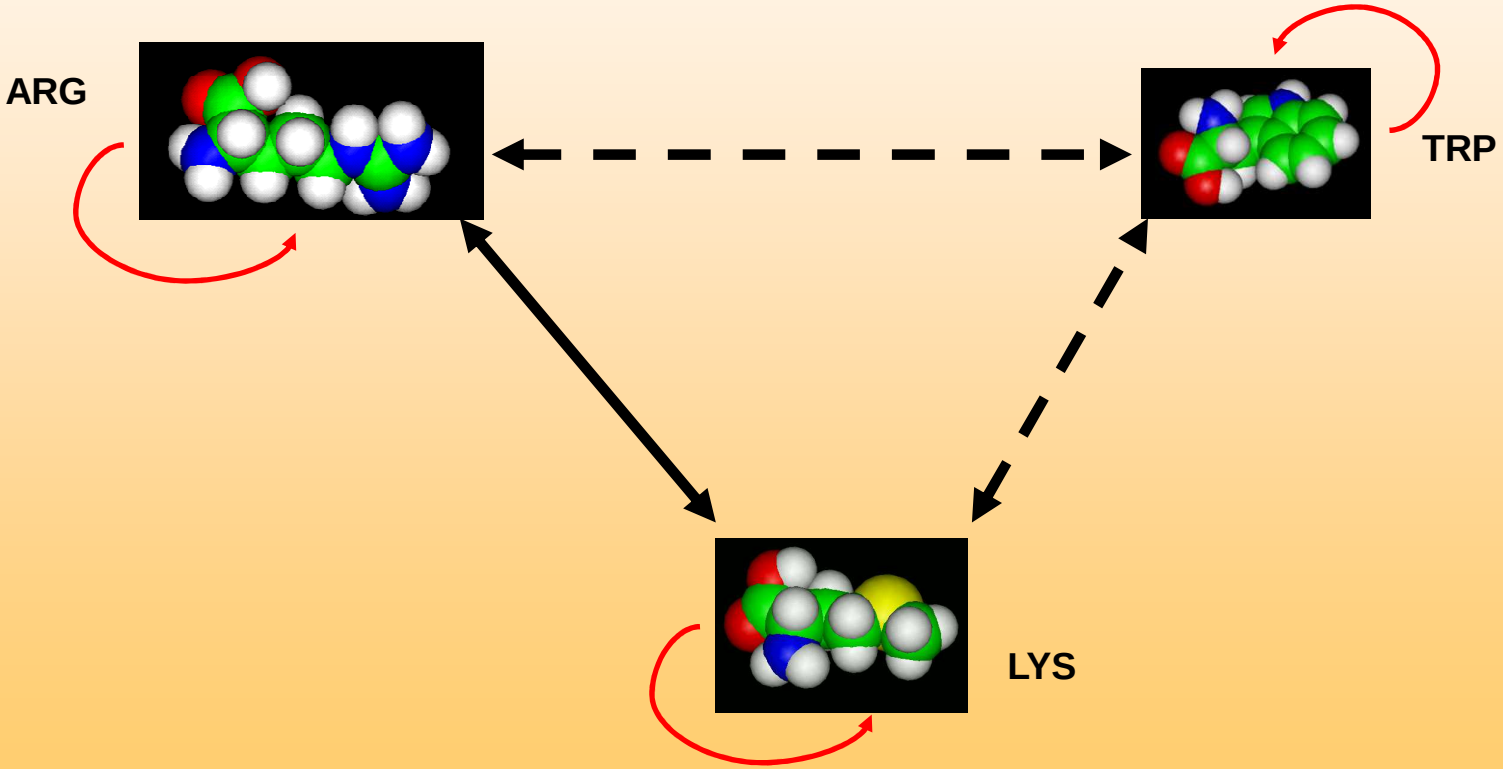


Reformulation du postulat dans le cadre de la théorie de l'information

- Les séquences de deux molécules de fonctions apparentées vont en général présenter une *information mutuelle* positive importante
- Réciproquement, deux molécules dont les séquences présentent une *information mutuelle* positive importante ont probablement des fonctions apparentées

La conservation de l'information mutuelle est incompatible avec la notion de distance

- Construire une distance $d(.,.)$ entre acides aminés (et donc entre séquences) nécessite la condition $d(a,a) = d(b,b) = 0$



Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

- Conclusion générale

Le CSHP, un espace abstrait

- Le CSHP: l'espace de configuration des protéines homologues, ou espace des séquences.

- Ne peut être appréhendé qu'à travers l'espace relatif à un référentiel, le $CSHP_{aref}$, avec a_{ref}, la séquence référence.

- Pour chaque séquence $b=b_1,\ldots,b_n$, ses coordonnées dans le $CSHP_{aref}$ sont les informations mutelles $I(a_i;b_i)$

- Pour un ensemble de x séquences, il est donc possible de considérer x $CSHP_{aref}$, chacun contenant une partie de l'information mutuelle totale du CHSP.

=====> espace de grande dimension dont:

1- le contenu est indissociable du contenant

2- les seules mesures disponibles sont les informations mutuelles totales et partielles du système

Une notion de proximité conservant l'information mutuelle: la q-dissimilarité (1)

Définition de la q-dissimilarité

On appelle *q-dissimilarité* dans E toute fonction $q(x,y): E \times E \to \Re^+$ telle que :

$$\text{i)} \quad \forall x \in E, \forall y \in E, q(x,x) = \min_{y \in E}\left(q(x,y)\right)$$

$$\text{ii)} \quad \forall x \in E, \forall y \in E, q(x,y) = q(y,x)$$

exp(-*s*) est une *q-dissimilarité*

Une nouvelle notion de proximité adaptée à la comparaison de séquence: la q-dissimilarité (2)

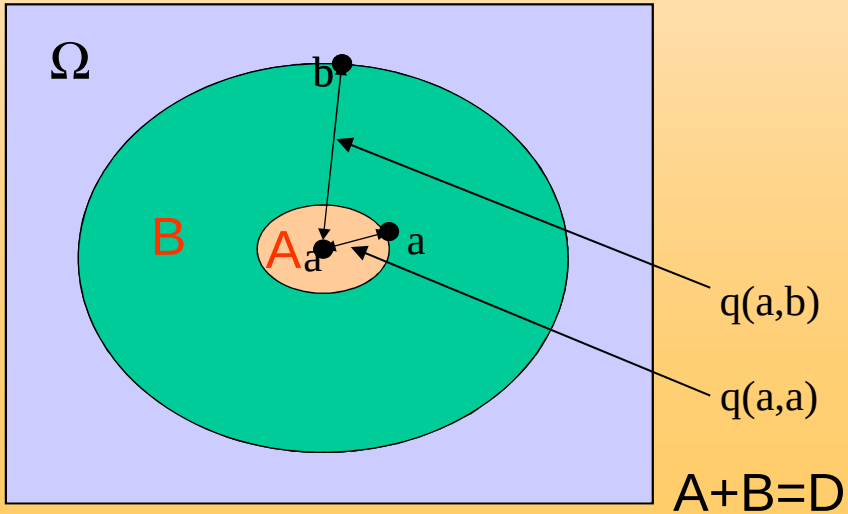
Corollaire

Avec le corollaire 2 de TULIP, on peut alors écrire :

$$P\left(Q(a,b^*) \leq q(a,b)\right) \leq \frac{1}{z(a,b^*)^2}$$

$$z(a,b^*) = \frac{s(a,b) - E[S(a,b^*)]}{\sigma[S(a,b^*)]}$$

Q(a,b*) la variable aléatoire quasi-dissimilarité de a et « b randomisé »



Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*
- La *Z-value* et l'estimation de la significativité d'un score d'alignement
- Comparaison de séquences dans le cadre de la théorie de l'information
- Un espace des séquences (CSHP) conservant l'information
- Le CSHP permet le calcul de distances évolutives
- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence
- Comparaison de séquences dans le cadre de la théorie de la fiabilité
- Conclusion générale

Modèle de la p-distance

- La distance évolutive, ou temps de divergence, entre 2 séquences est définie comme étant une fonction du nombre d'événement mutationnel (e.m.) par site sous tendant l'histoire évolutive de ces deux séquences

- Par définition, la p-distance, est égale à

$$pdist = 1 - y(a,b)$$

, y est le pourcentage de résidus identiques entre les 2 séquences

Exemple:

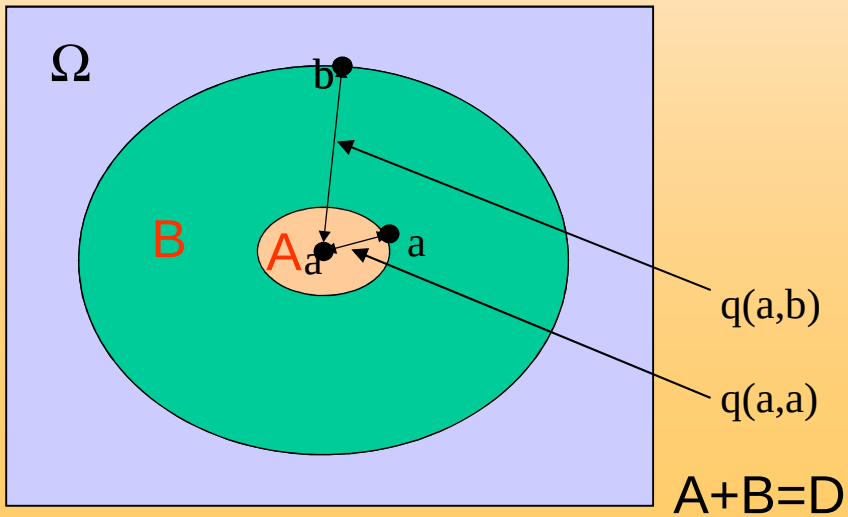
$$t(a,b) = -\log(y(a,b)) \qquad , y(a,b) = \frac{S(a,b) - S_{rand}(a,b)}{S(id) - S_{rand}(id)}$$

Une nouvelle approche probabiliste (1)

• $y(a,b)$ peut être interprété comme la probabilité que b partage la même information que a, connaissant au moins celle de b

• On peut alors définir:

$$y(a,b)=P\{Q(a,b^*)\leq q(a,a)/Q(a,b^*)\leq q(a,b)\}$$

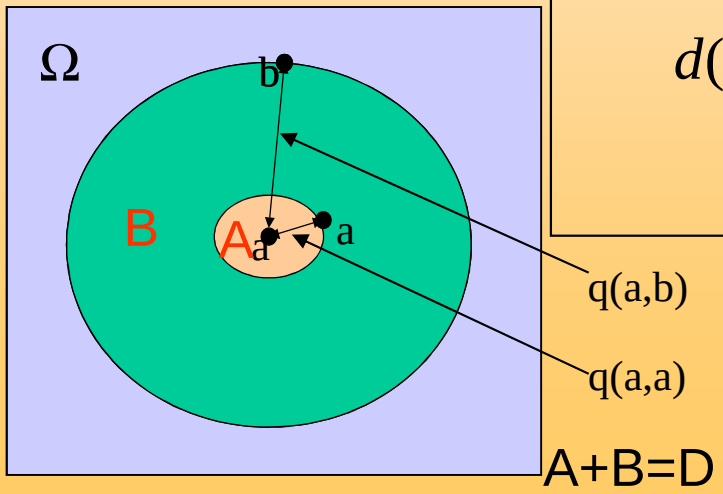


Une nouvelle approche probabiliste (2)

$$y(a,b)=\frac{P\{Q(a,b^*)\le q(a,a)\}}{P\{Q(a,b^*)\le q(a,b)\}}\le\frac{z^2(a,b^*)}{z^2(a,a^*)}$$

ANALOGIE AVEC LE MODELE DE FITCH:

$$d(a,b)=-\log(y(a,b)) \quad ,y(a,b)=\frac{S(a,b)-S_{rand}(a,b)}{S(id)-S_{rand}(id)}$$



Une nouvelle approche probabiliste (3)

Prise en compte des deux origines:

$$t_a(a,b) = -\log\left(\frac{z^2(a,b^*)}{z^2(a,a)}\right)$$

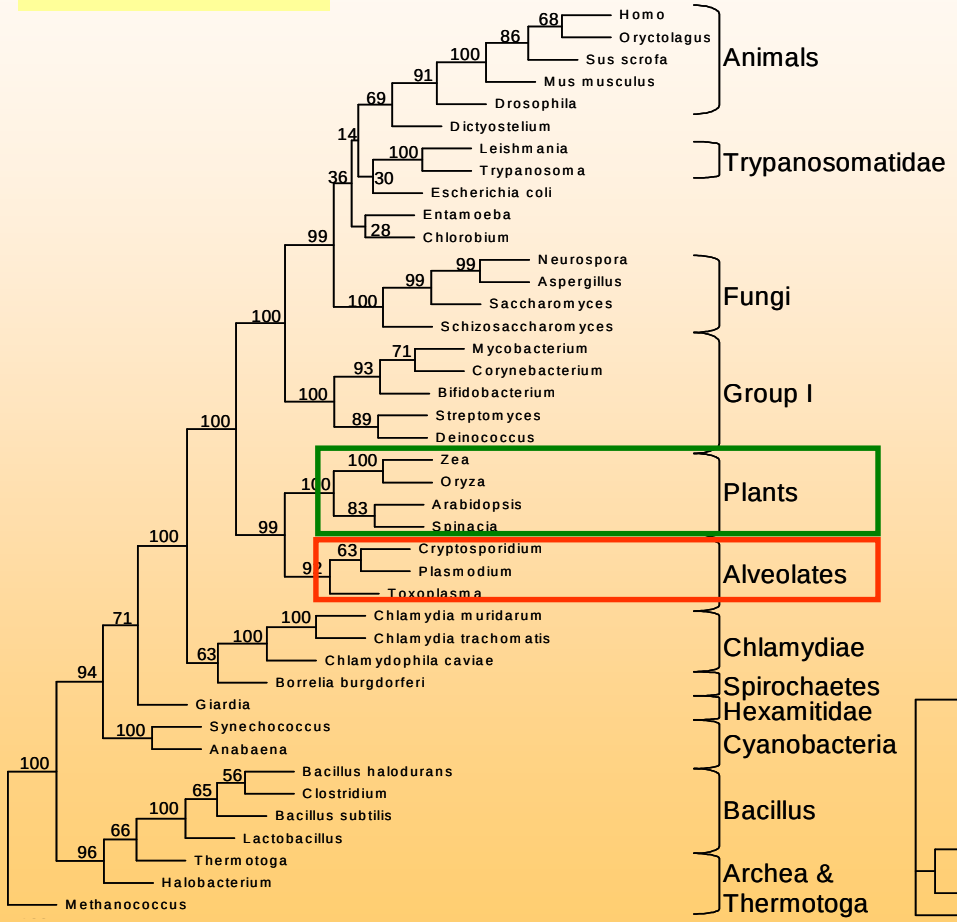
$$t_b(a,b) = -\log\left(\frac{z^2(b^*,a)}{z^2(b,b)}\right)$$

Calcul final de la distance évolutive:

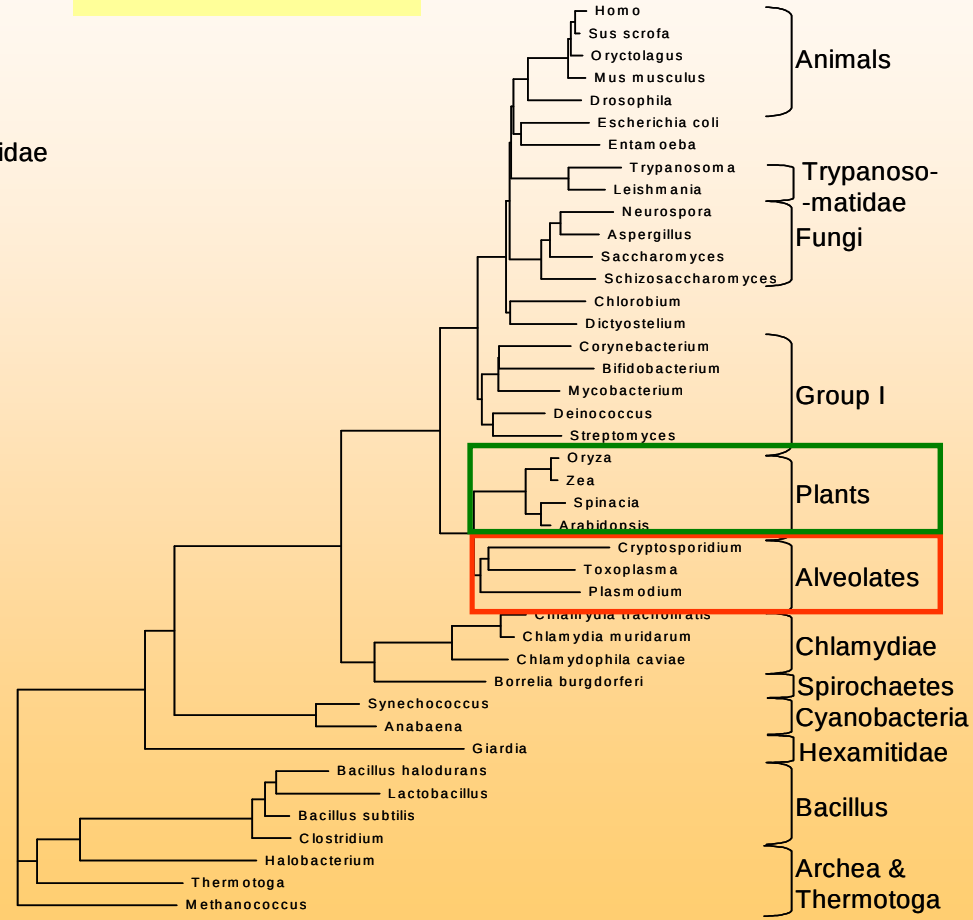
$$t(a,b) = (t_a^2 + t_b^2)^{1/2}$$

Exemple 1: Glucose-6-Phosphate Isomerase

MAB TREE

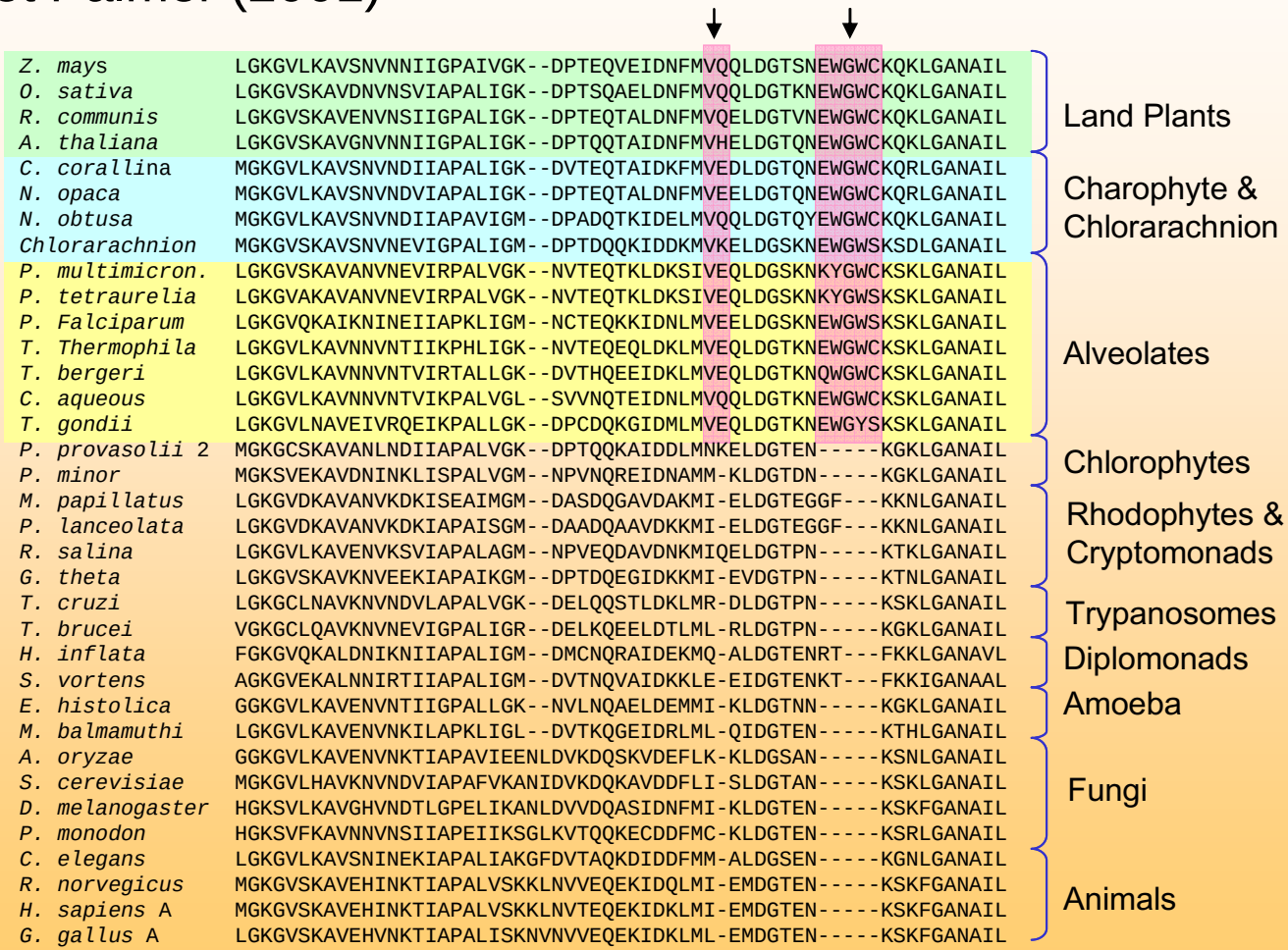


TULIP TREE



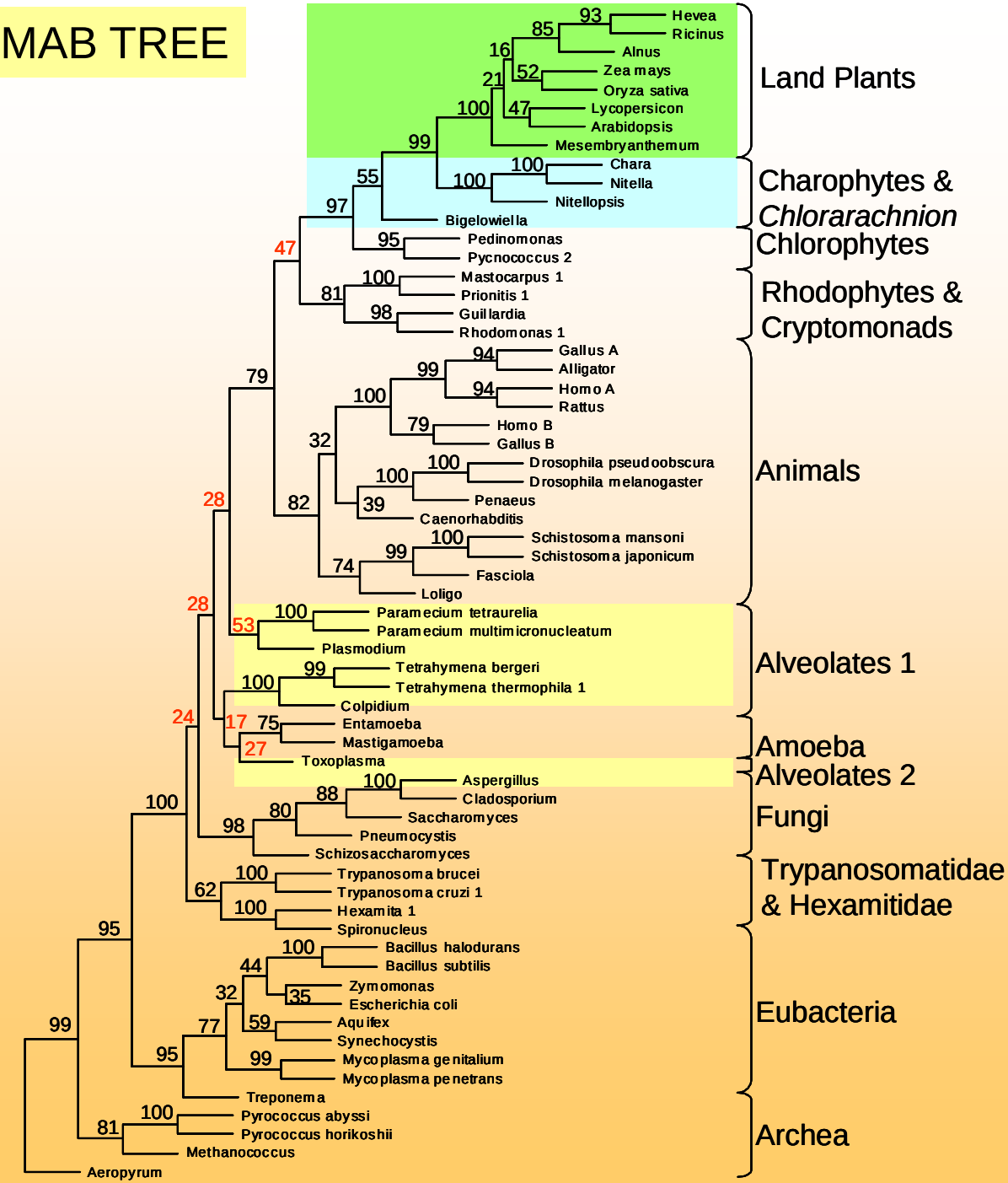
Exemple 2: l'énolase(1)

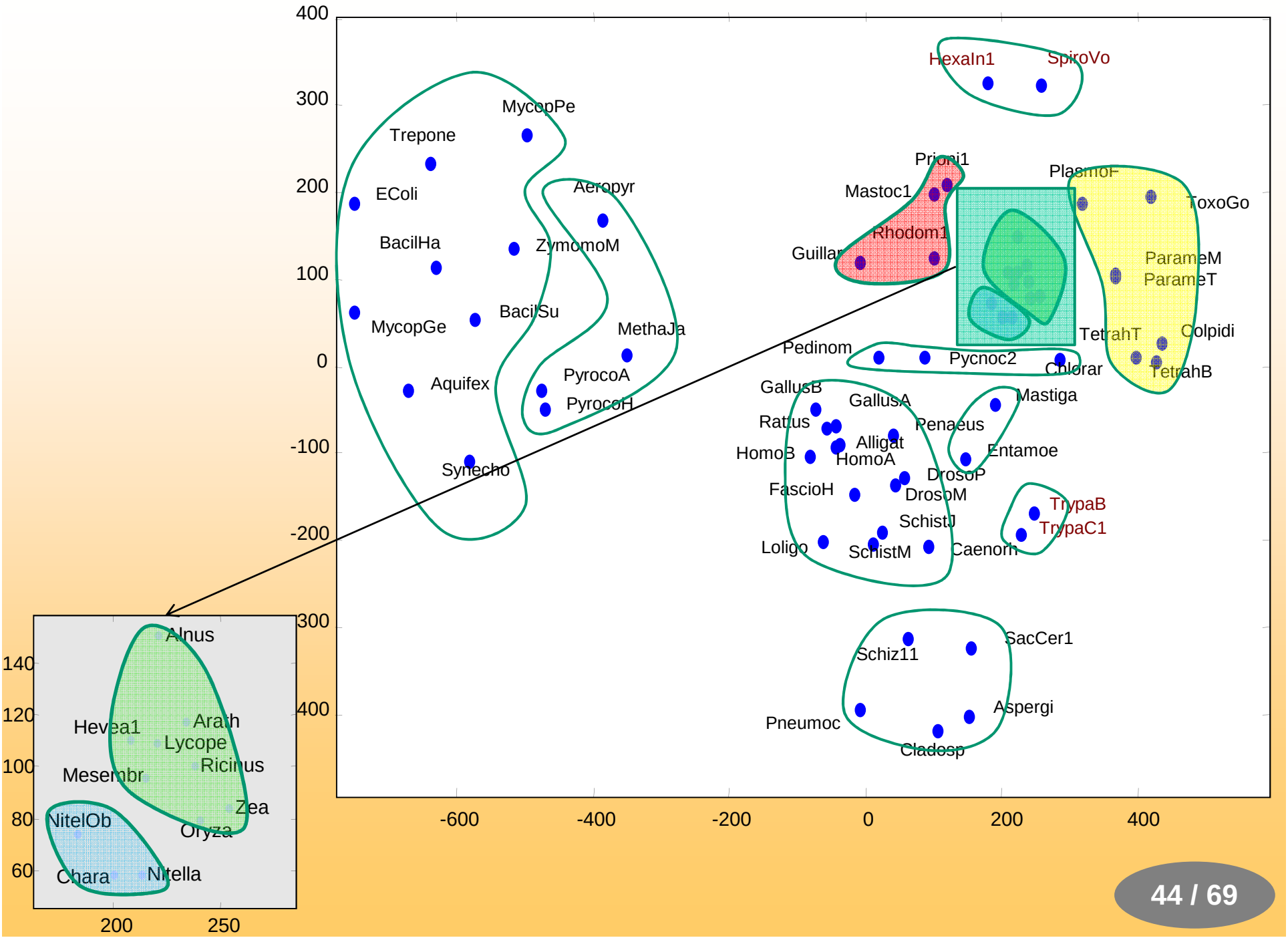
Keeling et Palmer (2001)



MAB TREE

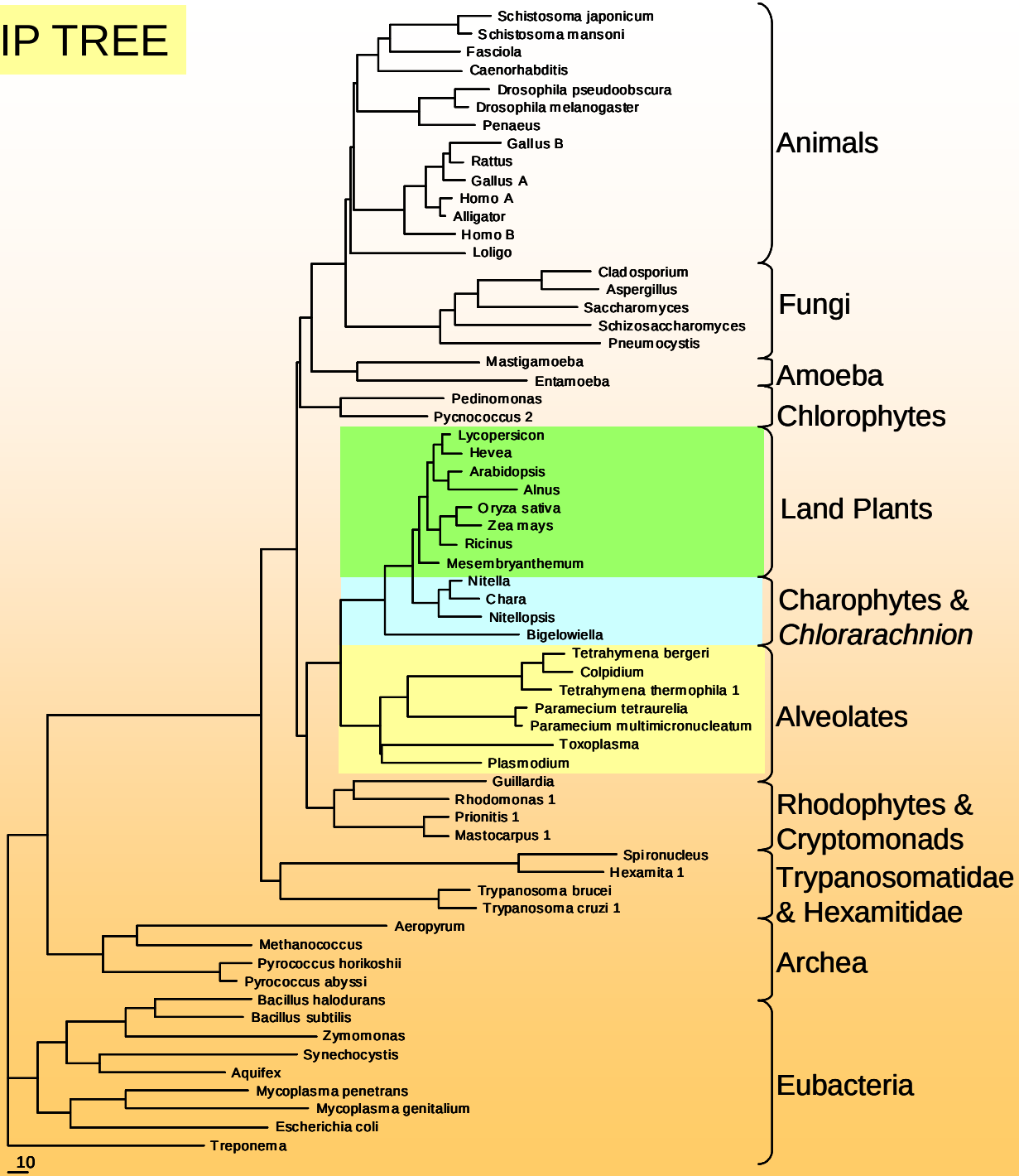
Exemple 2: l'énolase(2)




HexaIn1
SpiroVo
MycopPe
Trepone
EColi
Aeropyr
BacilHa
ZymomoM
MycopGe
BaciSu
MethaJa
Aquifex
PyrocoA
PyrocoH
Synecho
Prioni1
Mastoc1
Rhodom1
Guillar
PlasmoF
ToxoGo
ParameM
ParameT
TetrahT
Colpidi
TetrahB
Pedinom
Pycnoc2
Chlorar
GallusB
GallusA
Mastiga
Ratt:us
Penaeus
Alligat
HomoB
HomoA
Entamoe
DrosoP
FascioH
DrosoM
TrypaB
TrypaC1
SchistJ
Loligo
SchistM
Caenorh
SacCer1
Schiz11
Aspergi
Pneumoc
Cladosp
Alnus
Arath
Hevea1
Lycope
Ricinus
Mesembr
Zea
NitelOb
Oryza
Chara
Nitella
400
300
200
100
0
-100
-200
-300
-400
-600
-400
-200
0
200
400
140
120
100
80
60
200
250

TULIP TREE

Exemple 2: l'énolase(3)



Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

- Conclusion générale

Objectifs

- Estimer l'influence du biais en acide nucléique (a.n.) sur la composition en acides aminées (a.a.) des protéines chez *Plasmodium*

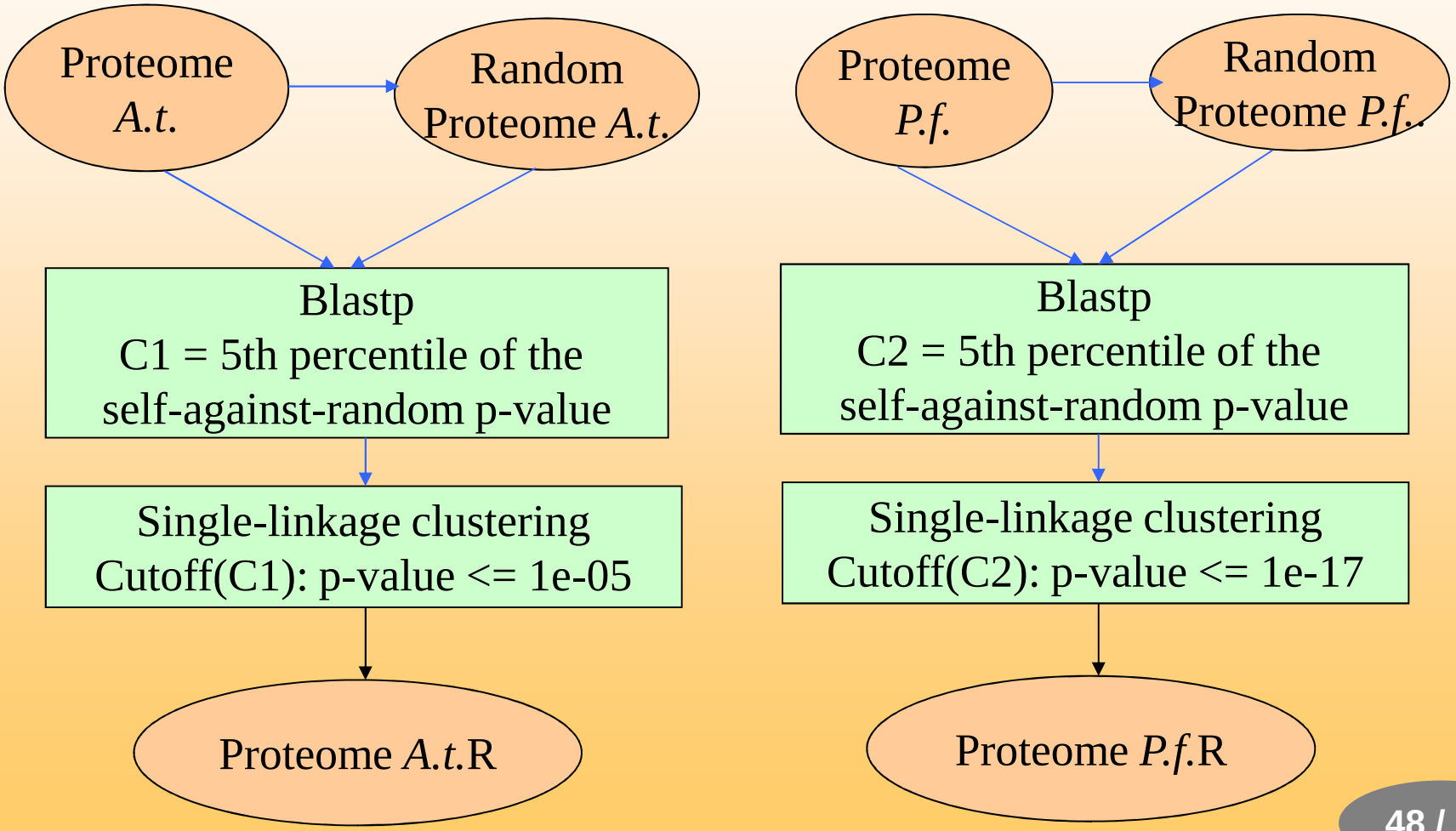
- Estimer la corrélation entre la composition (a.n., a.a.) des gènes et la pression de sélection s'excerçant sur leurs produits (les protéines)

Idée générale

- Utilisation du génome et du protéome d'*Arabidopsis* comme référence.

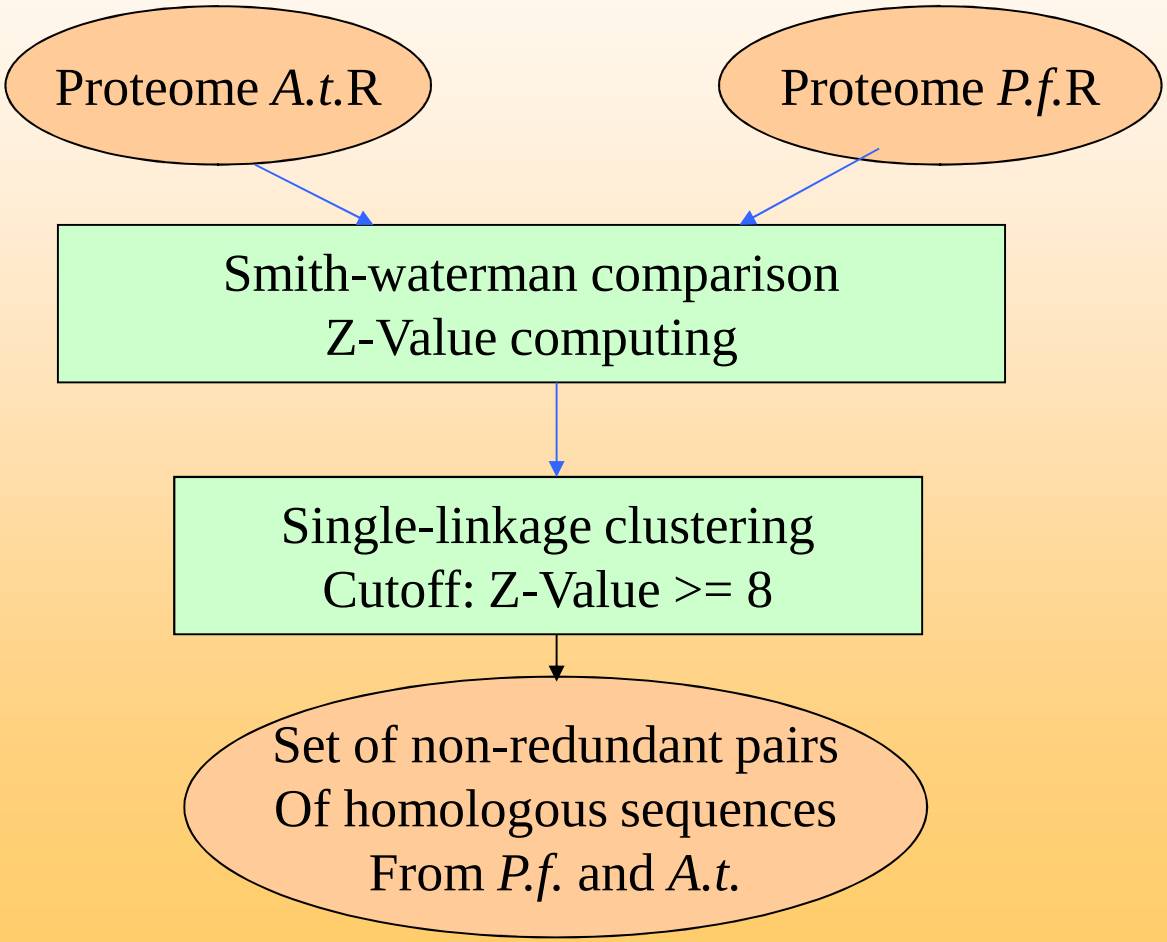
Comparaison des protéomes (1)

détermination des protéomes représentatifs, "R"



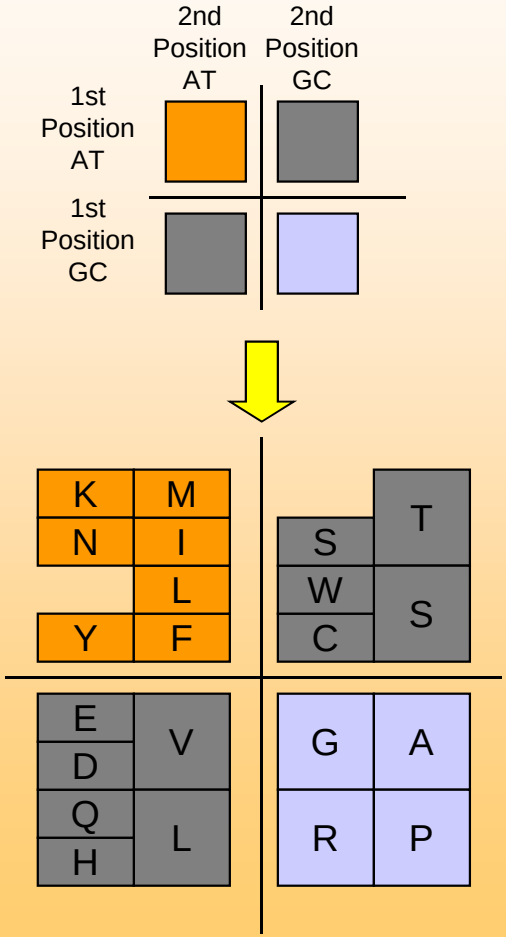
Comparaison des protéomes (2)

comparaison des protéomes R



Comparaison des protéomes (3)
Partitionnement de la table des codons

Foster et al, 1997



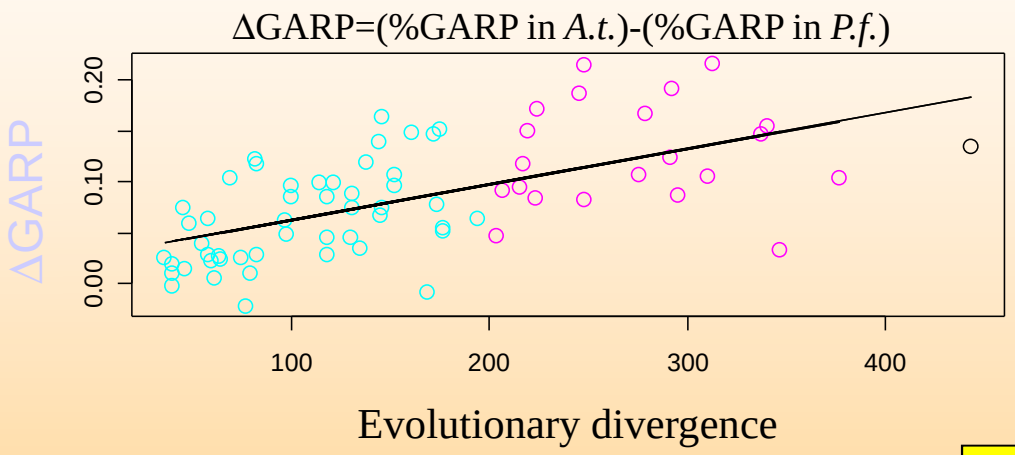
Hypothèses

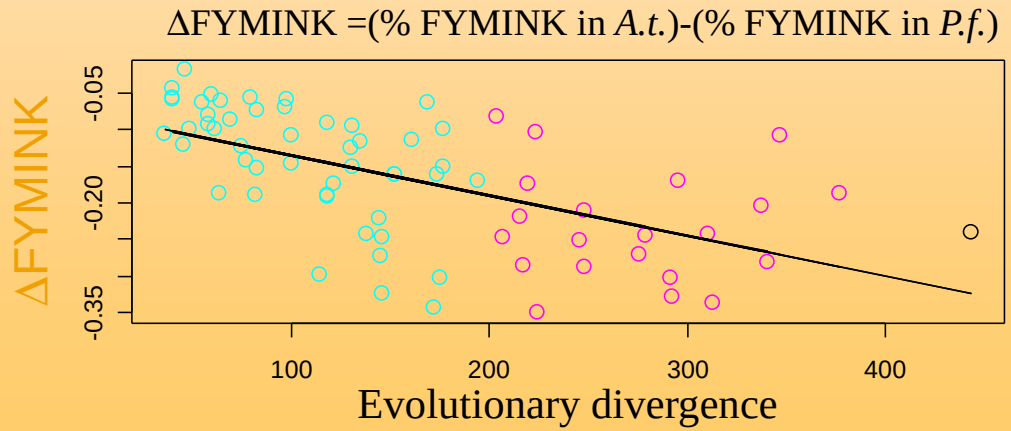
1- Les séquences riches en A+T vont coder pour des protéines enrichies en **FYMINK** par rapport à leurs homologues (moins riches en A+T).

2- Les séquences riches en G+C vont coder pour des protéines enrichies en **GARP** par rapport à leurs homologues (moins riches en G+C).

Comparaison des protéomes (4)

Relation entre la composition en aminoacides de séquences homologues et le degrés de divergence de ces séquences



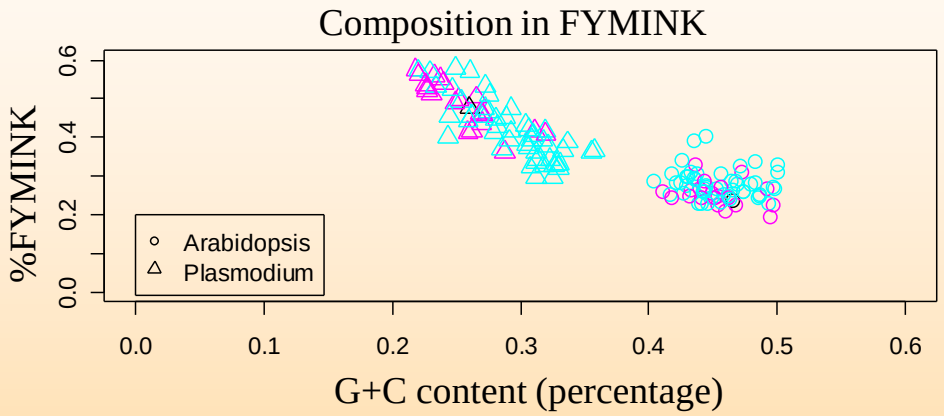


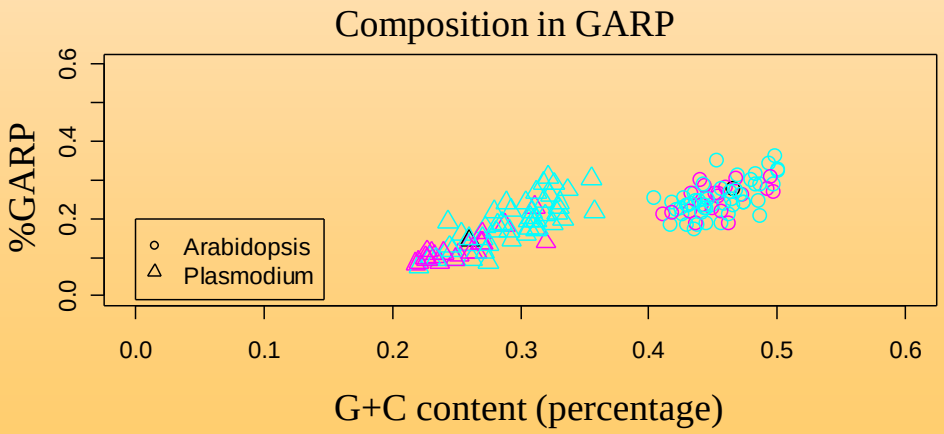
-Les protéomes de *A.t.* et *P.f.* ont des évolutions directionelles différentes.

- Les divergences de compositions (ΔGARP et ΔFYMINK) entre *A.t.* et *P.f.* sont majoritairement le fait de l'évolution de composition en a.a. des séquences de *P.f.* (pas de corrélation entre la composition des séquences d'*A.t.* et le temps évolutif).

Comparaison *A.t./P.f.*

Relation entre la composition en aminoacides et la composition G+C du CDS correspondant pour chaque paire de séquences homologues de *A.t.* and *P.f.*



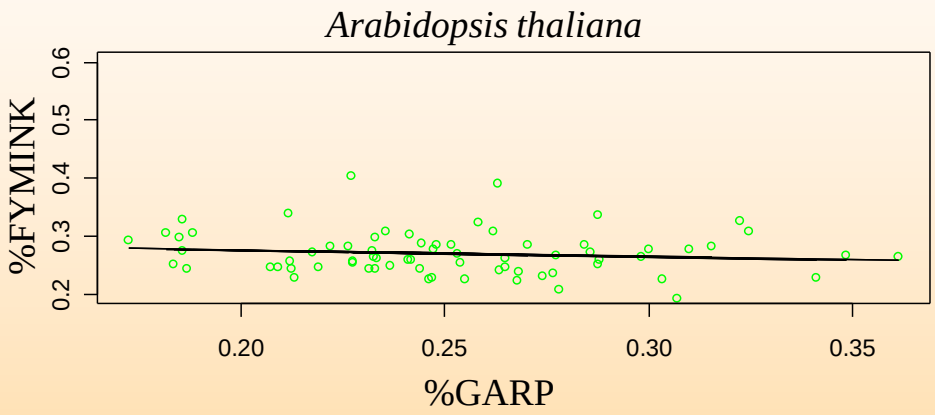


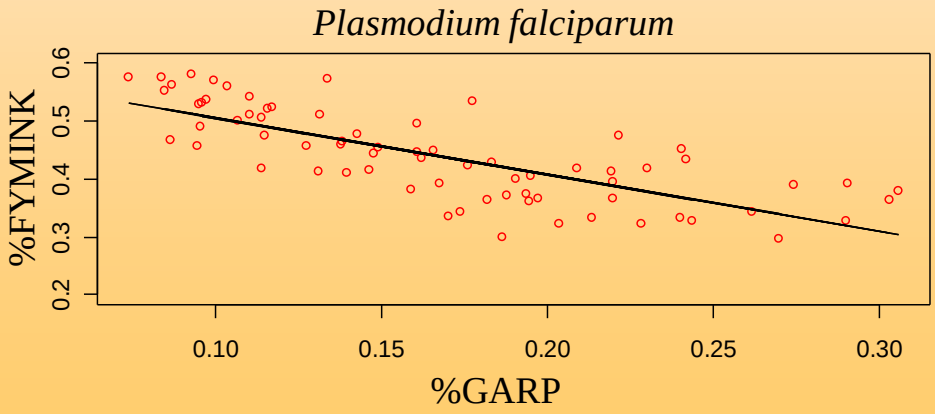
-Pas de recouvrement entre les compositions en G+C et en %GARP (ou %FYMINK) entre les 71 couples.

- Chez *P.f.*, le biais en A+T est plus fort dans la position synonyme des codons (81% contre 72% dans les positions non-synonymes)

Comparaison *A.t./P.f.*

Relation entre le %FYMINK et le %GARP pour chaque séquence de *A.t.* et *P.f.*





- Les acides aminés GARP sont substitués avec des acides aminés FYMINK dans le protéome de *Plasmodium falciparum*

Les matrices dirAtPf prennent en compte la différence de composition en aminoacides des protéines entre *A.t.* et *P.f.*

j dans la séquence **requête (*A.t.*)**
aligné avec ***k*** dans la séquence **sujette (*P.f.*)**,
avec une fréquence q_{jk}

$$dirAtPf(j,k)=\lambda.\log\left(\frac{q_{j,k}}{\pi_j\tau_k}\right)$$

Les matrices dirAtPf sont asymétriques

	A	R	N	D	C
A	5	-3	-2	-3	0
R	-1	6	-4	-1	-2
N	-1	0	4	1	-1
D	-2	-3	0	5	-4
C	1	-3	-2	-4	8

Arabidopsis thaliana

Plasmodium falciparum

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-3	-2	-3	0	-1	-3	0	-2	-2	-2	-4	-2	-3	-1	0	0	-1	-3	0
R	-1	6	-4	-1	-2	0	-2	-3	0	-3	-5	1	0	-7	-1	-1	0	-3	-5	-3
N	-1	0	4	1	-1	0	0	0	0	-3	-2	0	-2	-3	0	0	0	-2	-2	-2
D	-2	-3	0	5	-4	0	1	-1	0	-3	-5	0	-1	-5	0	0	-1	-6	-5	-3
C	1	-3	-2	-4	8	-3	-5	0	-2	-1	0	-5	-1	-2	-2	0	0	0	0	0
Q	0	0	0	-1	0	5	0	-2	0	-5	-1	0	0	-6	0	0	-2	0	0	-3
E	0	0	0	1	-3	1	4	-1	-1	-4	-3	0	-2	-5	0	0	0	0	-4	-2
G	0	-3	0	-6	-2	-4	-4	7	-1	-3	-5	-4	-4	-5	-3	-2	-2	-5	-4	-5
H	-2	1	0	-1	-4	1	0	-2	7	-4	-2	-3	-2	-1	-1	0	-1	0	1	-2
I	-1	-3	-3	-4	0	-2	-2	-4	-3	4	1	-2	0	0	-1	-2	0	0	-2	2
L	-1	-2	-4	-4	0	-2	-3	-4	-2	1	4	-3	2	0	-5	-2	-1	0	-1	0
K	0	1	0	0	-4	0	0	-2	0	-3	-2	4	0	-5	0	0	0	-5	-2	-2
M	0	-1	-2	-2	0	-2	-4	-3	0	0	2	-2	6	0	-1	-1	-1	-4	-3	0
F	-1	-1	-3	-2	-1	-3	-3	-3	-1	0	0	-4	0	6	-3	-3	-4	1	2	0
P	-1	-2	-1	-1	-2	-1	-1	-5	-1	-7	-3	-1	-4	-1	7	-1	-1	-4	-2	-3
S	1	-2	0	-1	0	0	0	0	-1	-3	-5	0	-2	-2	0	4	0	-3	-3	-3
T	0	0	-1	-1	-1	0	0	-1	-1	-1	-1	-1	0	-2	-1	1	5	-4	-5	0
W	-6	-1	-5	-3	0	-4	-4	-3	-1	-2	0	-3	0	3	-5	-3	-3	11	2	-4
Y	-1	0	-2	-2	-1	-1	-1	-2	1	-2	0	-2	0	3	-2	-1	-1	3	6	-1
V	0	-3	-5	-7	0	-3	-2	-2	-4	2	0	-2	0	0	-2	-3	0	-2	-2	4

Les matrices dirAtPf validées pour leur usage dans les recherches d ’homologues

- Gain théorique en sensibilité
  Entropie relative haute

- Gain en spécificité validé par l ’expérimentation
  Implémentation dans le moteur de recherche Blastp et dans smith-waterman:
    dirAtPf génère moins de faux positifs Blosum, Pam

Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

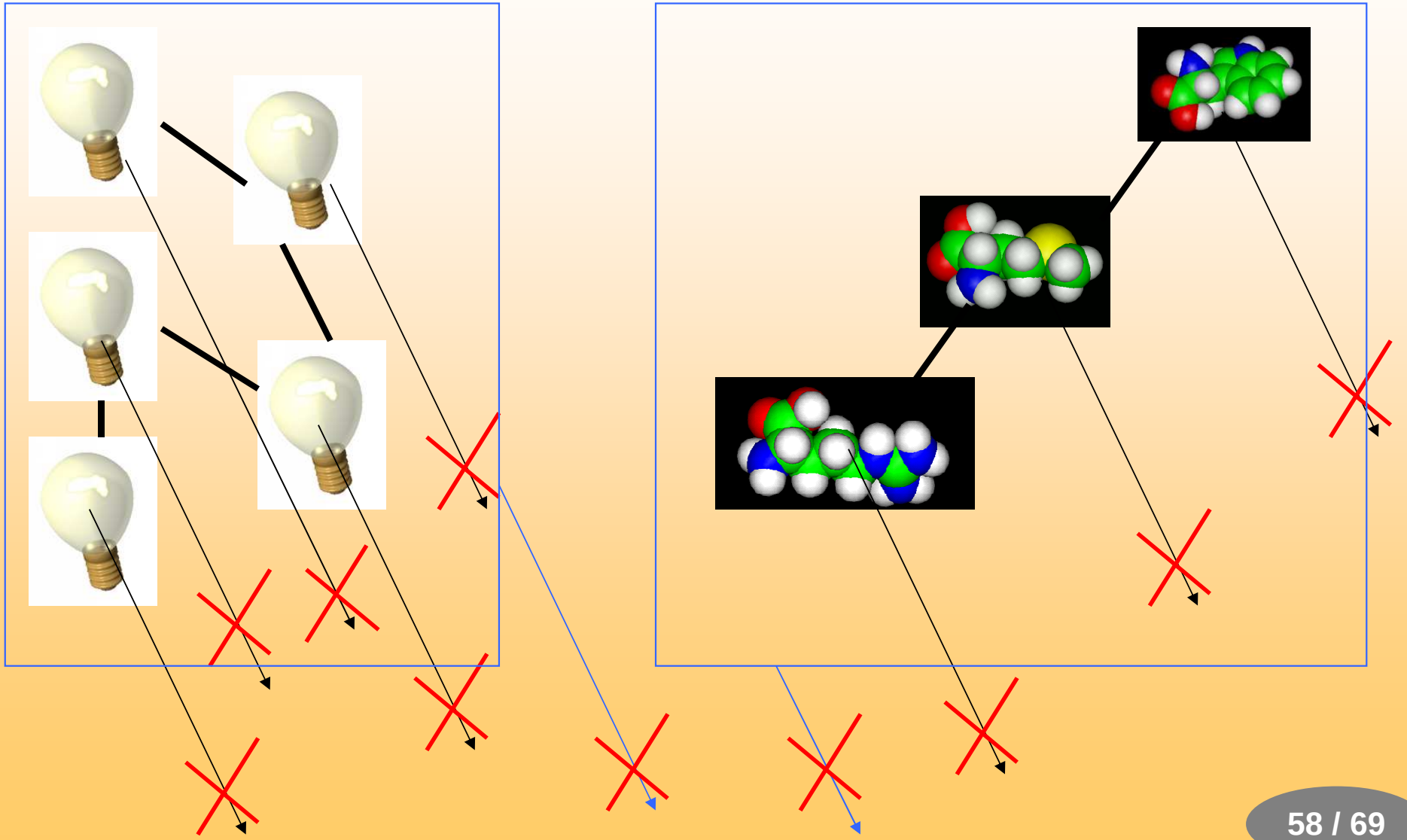
- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

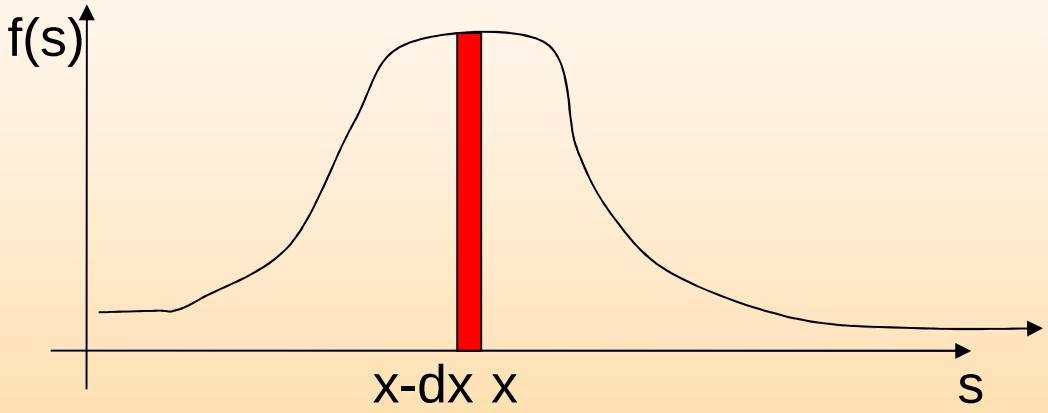
- Conclusion générale

La théorie de la fiabilité



La fonction de longévité

On se donne *la loi de probabilité* $P(X \leq x)$ (=$F(x)$) avec la *densité* correspondante $f(x)=dF/dx$.



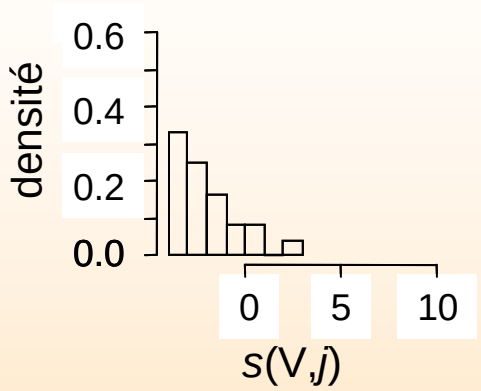
Fonction de longévité: Probabilité de mourir entre x-dx et x, sachant que cela se produit entre 0 et x

$$\psi(x)=\lim_{dx \to 0} \frac{P(x-dx < X \leq x / X \leq x)}{dx}$$

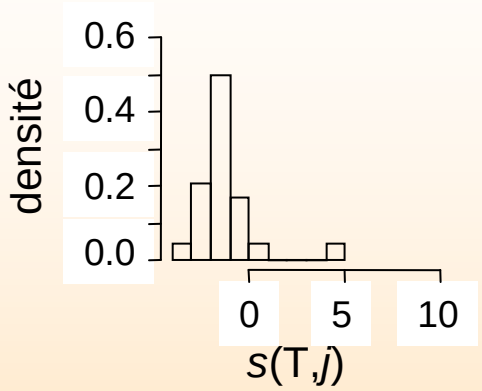
$$\psi(x)=\frac{f(x)}{F(x)}=\frac{f(x)}{P(x \leq X)}$$

L’évolution des monomères en fonction du temps

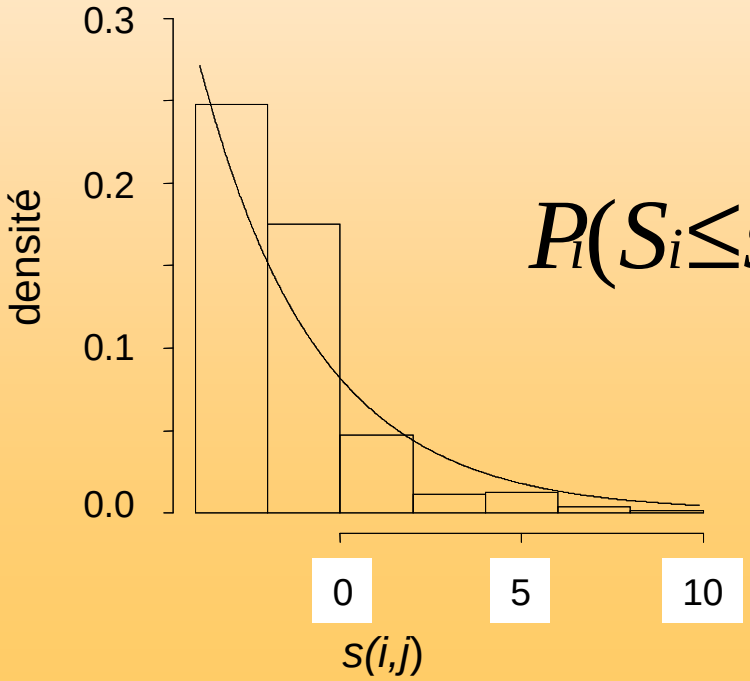
Valine: un composant non-vieillissant



Thréonine: un composant vieillissant



Tous les résidus (basé sur BLOSUM62): composants non-vieillissants



$$P_i(S_i \leq s_i) = 1 - \exp(-\lambda . s_i)$$

L’hypothèses d’homogénéité statistique de l’information mutuelle

- On pose s(a,b) le score observé entre a et b, S le score aléatoire entre les séquences aléatoires A et B

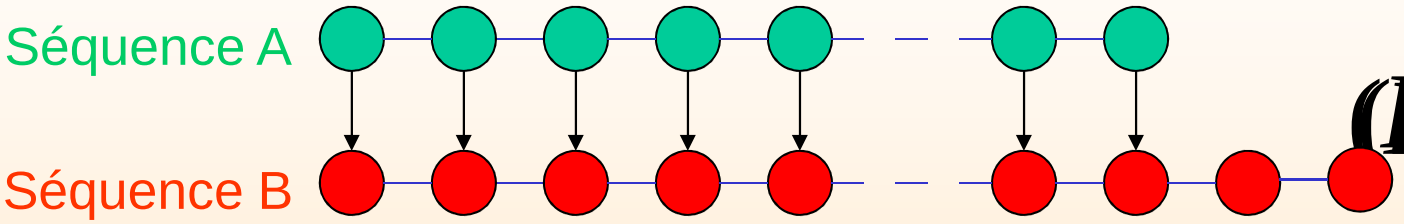
- On défini les approximations suivante, ou m est la taille de l’alignement entre a et b

$$S \approx m\langle S_i \rangle \quad \text{Où idéalement,} \quad \langle S_i \rangle = \lim_{m \to +\infty} \frac{S}{m}$$

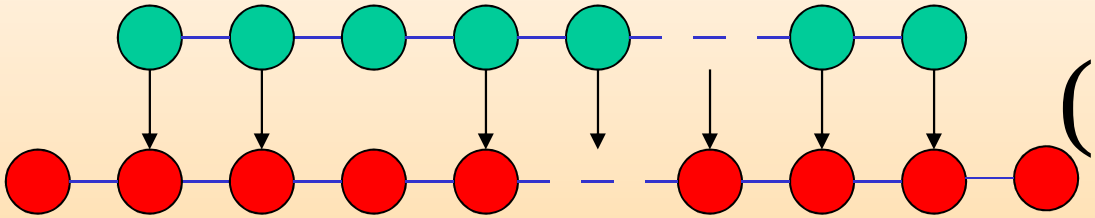
$$s \approx m\langle s_i \rangle \quad \text{Où idéalement,} \quad \langle s_i \rangle = \lim_{m \to +\infty} \frac{s}{m}$$

$$\Rightarrow P_i(S_i \leq s_i) \approx P_i(S \leq s)$$

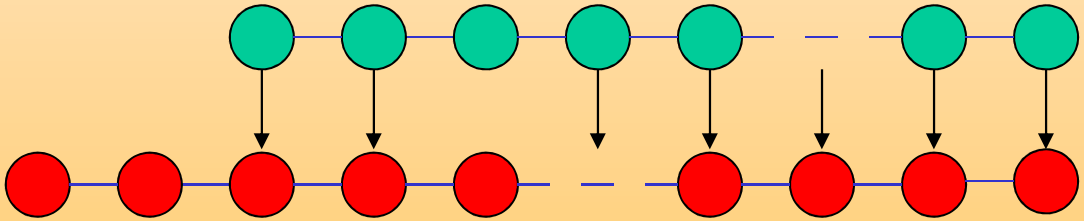
Le calcul de la loi de probabilité (1)



$$(P_i(S_i \le s_i))^m$$



$$((P_i(S \le s))^m)^2$$



$$((P_i(S \le s))^m)^{Km}$$

$$\Rightarrow P(S \le s) = (P_i(S \le s))^{K(a,b)mn}$$

Le calcul de la loi de probabilité (2)

De la forme de la loi de la probabilité, $P(S \leq s) = (P_1(S \leq s))^{K(a,b)mn}$

On déduit celle de la densité de probabilité et donc celle de la fonction de longévité

$$\psi(s) = \frac{K.m.n.\lambda.\exp(-\lambda s))}{1 - \exp(-\lambda s))}$$

Et donc asymptotiquement, $\psi(s) \approx K.m.n.\lambda.\exp(-\lambda s))$

Ce qui conduit permet de déduire la loi de probabilité ds scores d'alignements

$$P(S \leq s) = \exp(-K.m.n.e^{-\lambda s})$$

La loi de probabilité de la *Z-value* est indépendante de la composition et de la taille des séquences

Faisant le changement de variable $Z=\frac{s(a,b)-\mu}{\sigma}$ et en utilisant les relations de Gumbel $\mu = \theta + \gamma\beta$ et $\sigma^2=(\pi^2/6)\beta^2$, on obtiens la distribution de probabilité de la *Z-value*:

$$P(Z\leq z)=\exp(-\exp(-z\frac{\pi}{\sqrt{6}}-\gamma))$$

Comparaison des différentes statistiques de score d'alignements

Cas des Facteurs de Transcriptions TFIIA gamma

Alignment method	Blastp	Smith-Waterman	
Substitution matrix	BLOSUM62	BLOSUM62	DirAtPf100
Statistics			
P-value (Karlin-Altchul)	0.008	NA	NA
Z-value (Pearson-Lipman)	10	11	12
T-value (TULIP theorem)	0.01	8.10^{-3}	7.10^{-3}
P-value (this work)	$1.5.10^{-6}$	$3.7.10^{-7}$	1.10^{-7}

Plan

- Problématique générale
  - les carences du génome malarial
  - la face végétale de *Plasmodium falciparum*

- La *Z-value* et l'estimation de la significativité d'un score d'alignement

- Comparaison de séquences dans le cadre de la théorie de l'information

- Un espace des séquences (CSHP) conservant l'information

- Le CSHP permet le calcul de distances évolutives

- Analyse du biais de composition du génome et du protéome de *Plasmodium* en utilisant *Arabidopsis* comme référence

- Comparaison de séquences dans le cadre de la théorie de la fiabilité

- Conclusion générale

Conclusion générale (1)

- La théorie de l'information fournit un cadre adapté pour reformuler certains principes néo-Darwinien dans des termes mathématiques. L'alignement de séquences replacé dans son contexte scientifique (la recherche de relation biologique) a conduit à de nombreux résultats

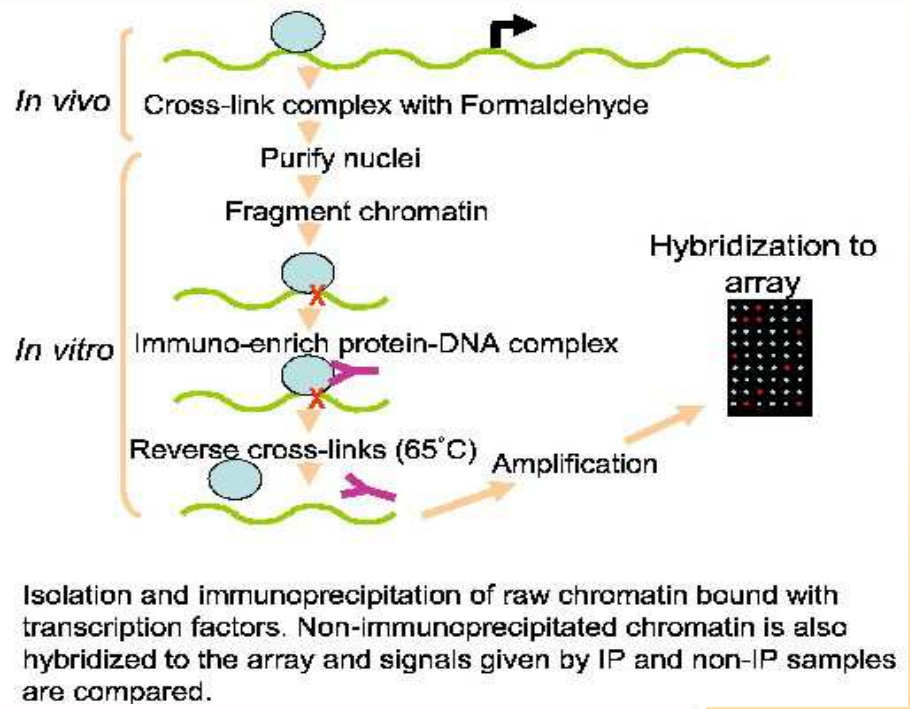
- Le modèle CSHP permet de construire des arbres phylogénétiques en tenant compte de la totalité de l'information mutuelle du système. Il permet également l'exploration de l'espace des protéines et donc la recherche de cibles herbicides comme voulu au départ du projet de thèse

Conclusion générale (2)

- Etude des interactions ADN/Protéines responsable de la régulation de l'état de la chromatine, incluant la régution dite siARN

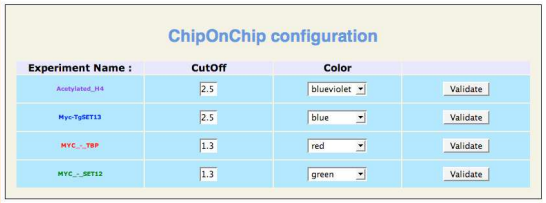
- Etude l'évolution des protéibes responsables de ses interactions (Domaines SET, …)

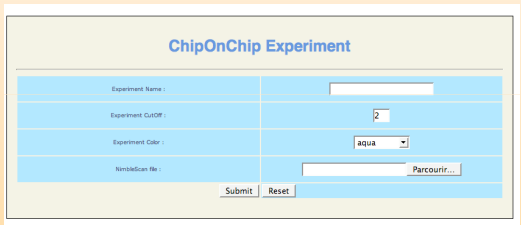
GenoBrowser :Le module *GenoChipOnChip*

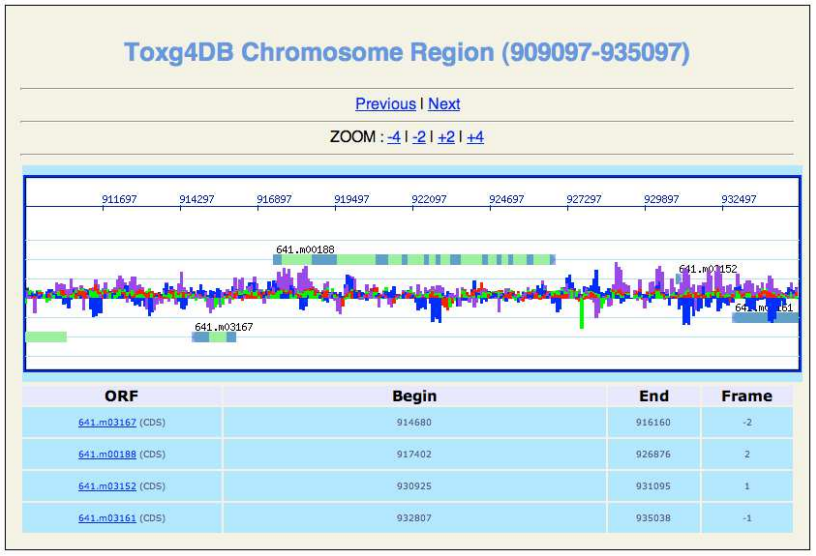


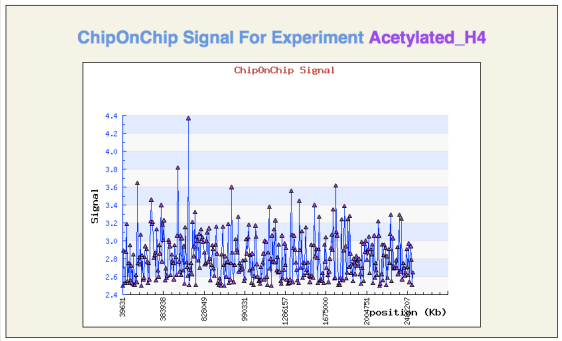
Isolation and immunoprecipitation of raw chromatin bound with transcription factors. Non-immunoprecipitated chromatin is also hybridized to the array and signals given by IP and non-IP samples are compared.

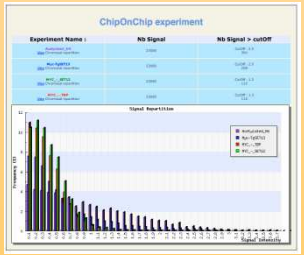
- ✓ Importation et configuration des expériences,
- ✓ Visualisation des résultats
  - ✓ global pour une expérience,
  - ✓ dans contexte génomique,



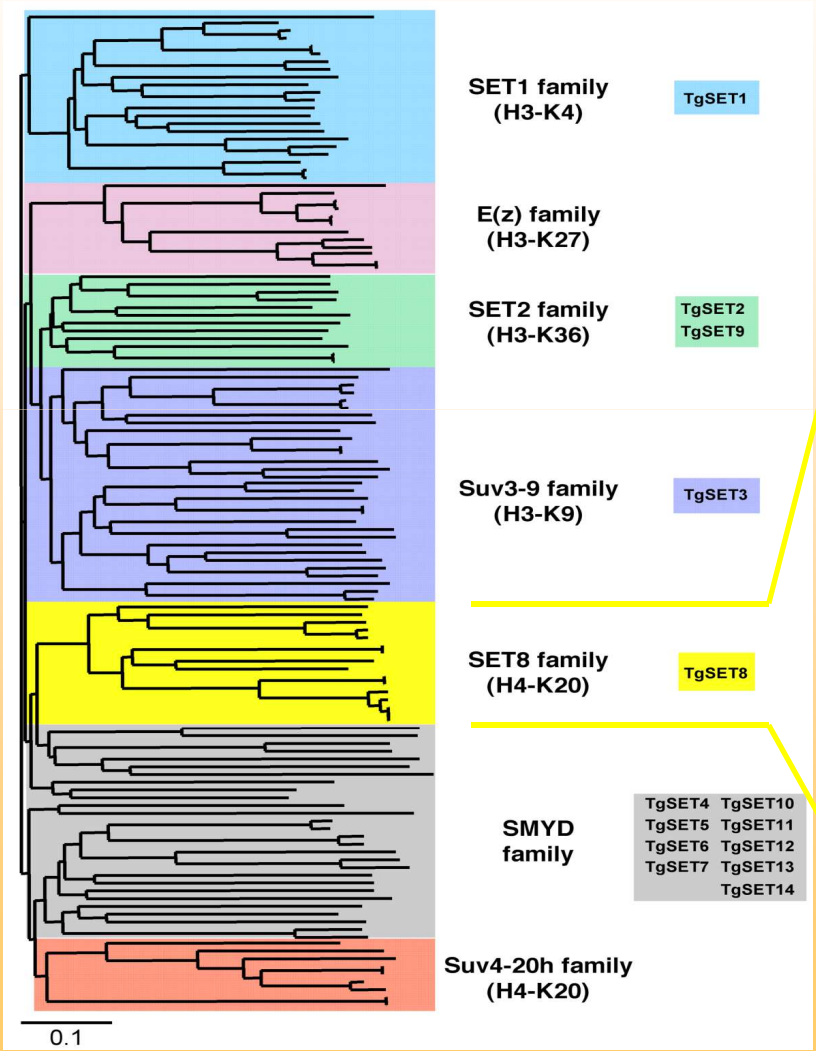


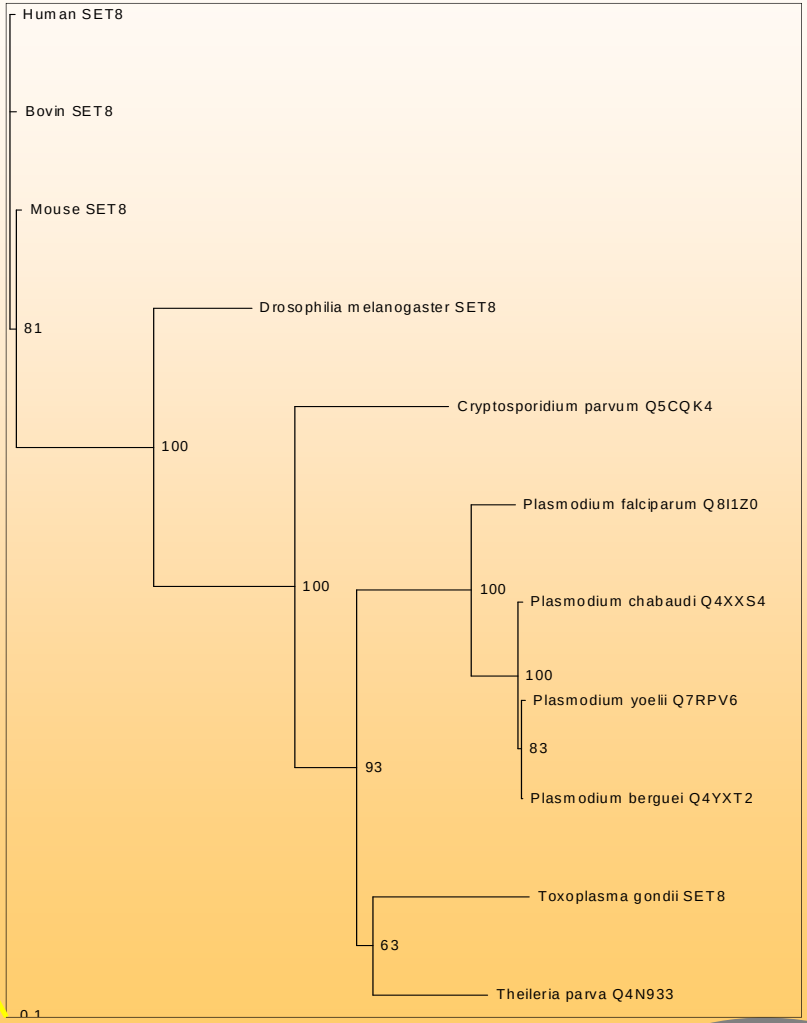






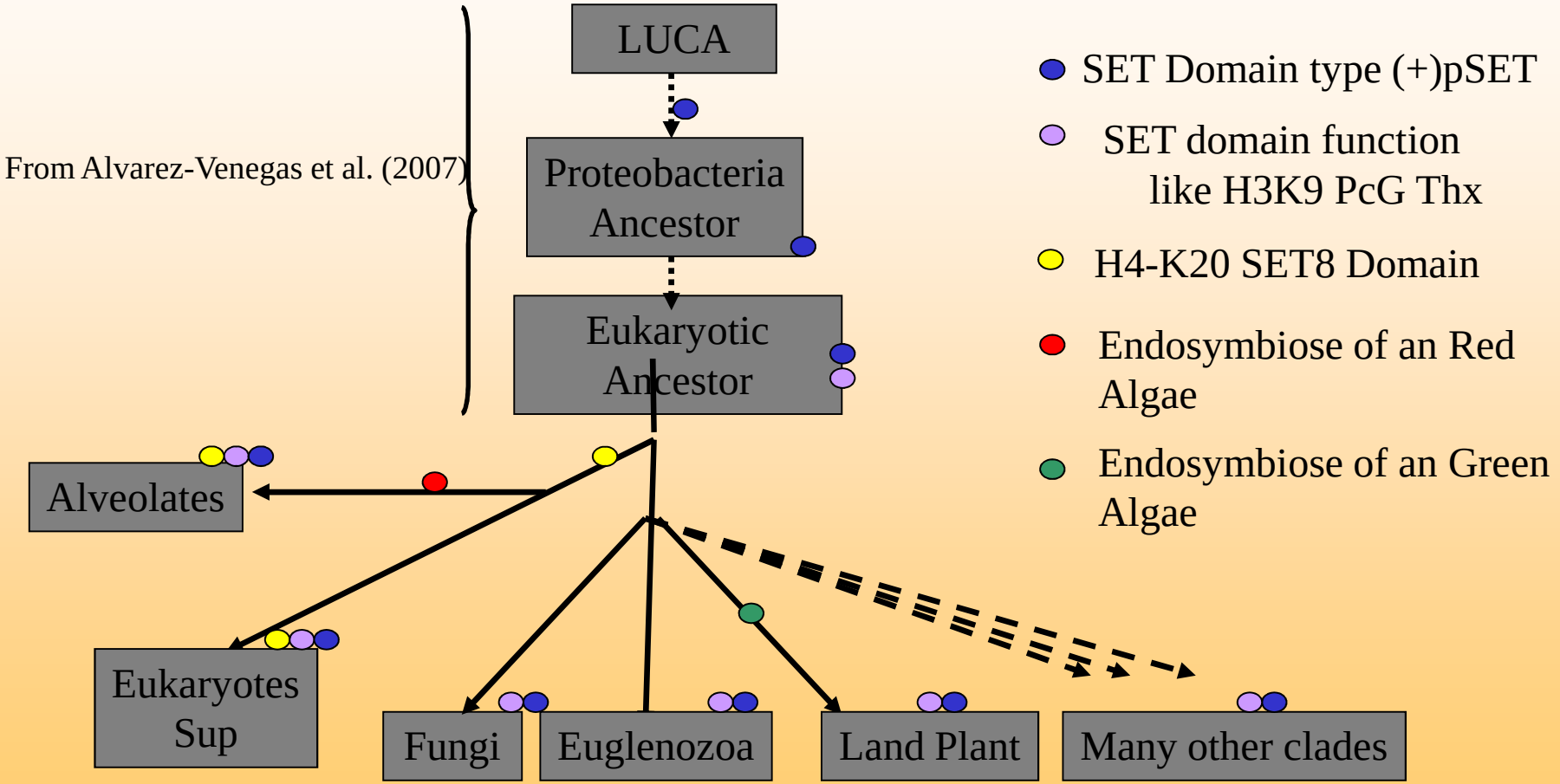
Identification of a New Family of HMTase Related to Human Set8 in Apicomplexan Parasites





A model for the existence of the H4-K20 SET domain in the Alveolates and Metazoan clades

The existence of two clusters (Alveolates and Metazoan) for the H4K20 SET domain in the phylogenetic tree of the SET domain phylogeny was support by the DD-HDS analysis.



Remerciements

Laboratoire de Physiologie Cellulaire Végétale (CEA Grenoble)
Eric Maréchal
Delphine Grando
Hélène Valadié

Laboratoire Biologie, Informatique et mathématiques (CEA Grenoble)
Sylvaine Roy

Gene-It
Jean-Jacques Codani
Karine Metayer

Laboratoire Imagerie Médicale quantitative (Hopital de la Pitié-Salpêtrière
Sylvain Lespinats
Bernard Fertil

Laboratoire de Bioinformatique, Génomique et Modélisation (CEA Saclay)
Jean-Christophe Aude

Département d'Écophysiologie Végétale et Microbienne (CEA Cadarache)
Philippe Ortet
Mohamed Barakat
Thierry Heulin

Laboratoire Adaptation et Pathogénie des Microorganismes
Mohamed-Ali Hakimi
Céline Sautel