Comparaison de génomes : Développement théoriques et méthodes numériques pour les analyses comparatives de génomes et protéomes biaisés

Application à la comparaison des génomes et protéomes de *Plasmodium falciparum* et d *'Arabidopsis thaliana*

Olivier Bastien

Plan

- Problématique générale

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence dans le cadre de la théorie de la fiabilité

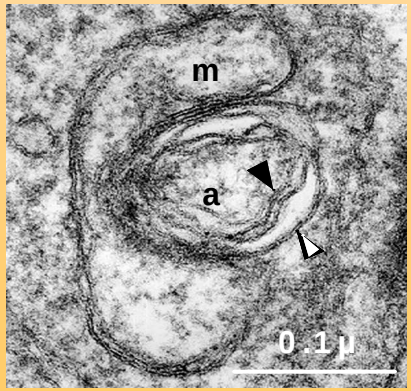
- Conclusion générale

Plasmodium falciparum:
l ’agent du paludisme

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

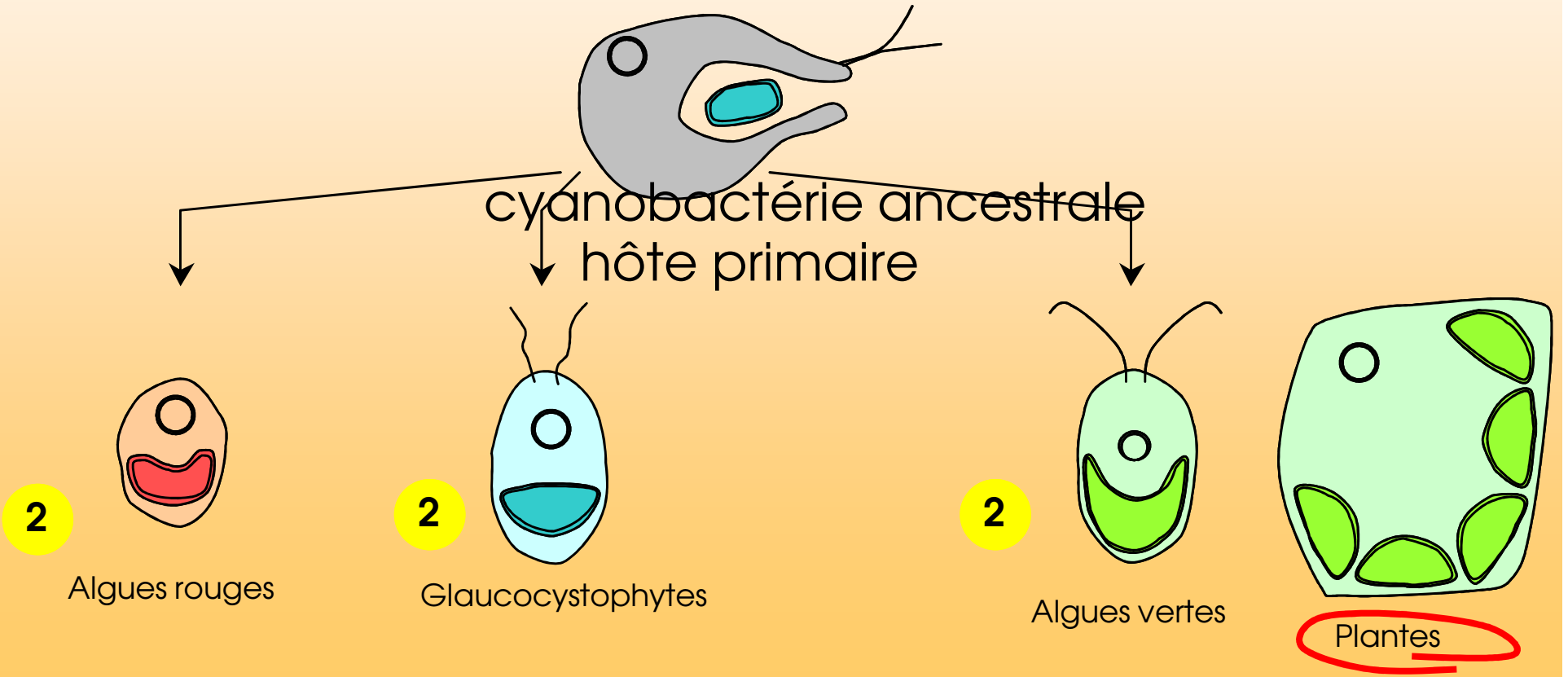
- Agent infectieux: *Plasmodium falciparum*
(→ séquençage complet depuis Octobre 2002)

- *Plasmodium falciparum* contient des structures typiquement végétales

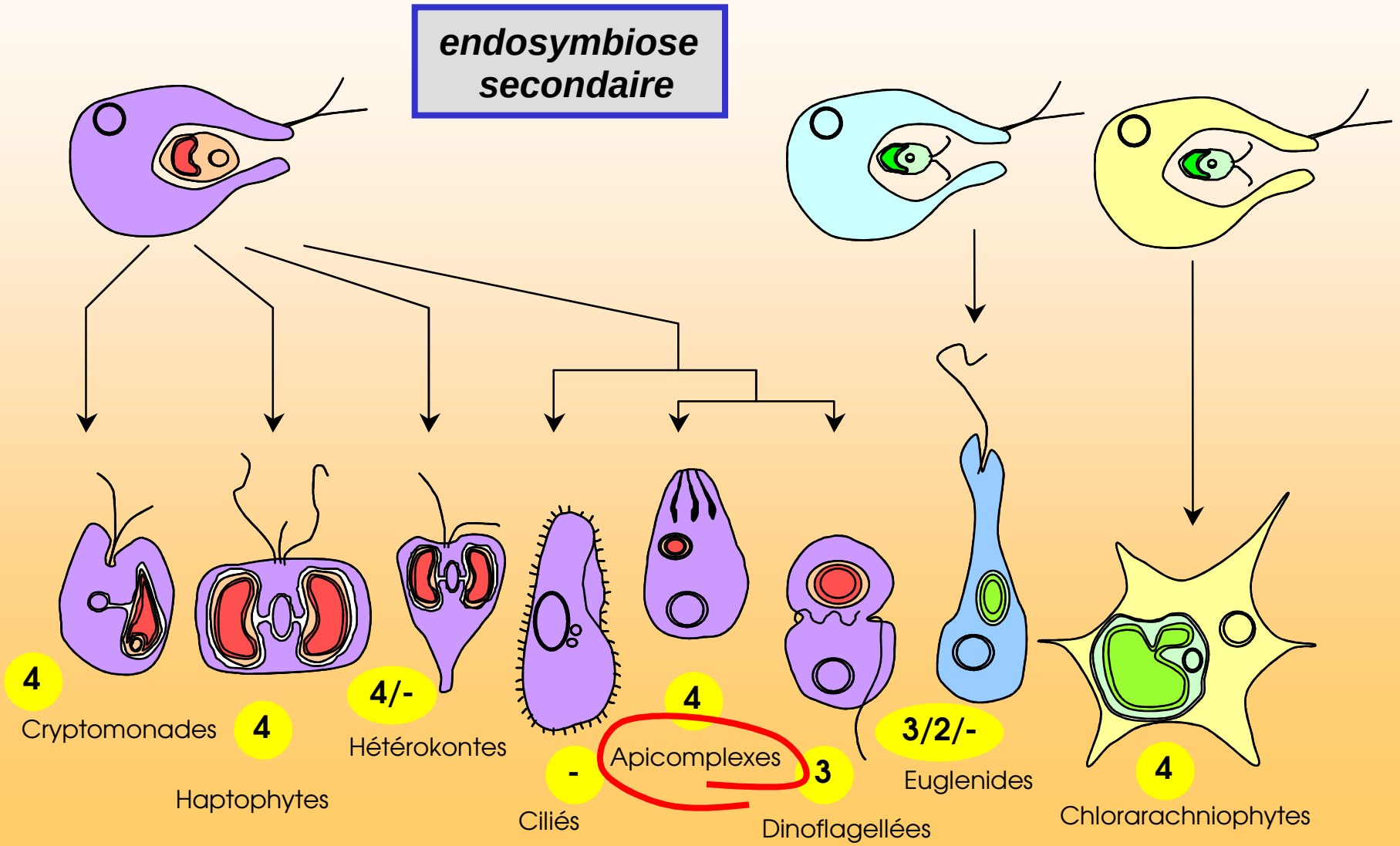


Plasmodium falciparum: Une histoire évolutive complexe (1)

endosymbiose primaire



Plasmodium falciparum: Une histoire évolutive complexe (2)



Plasmodium falciparum: *un génome nucléaire composite*

issus de la cyanobactérie ancestrale

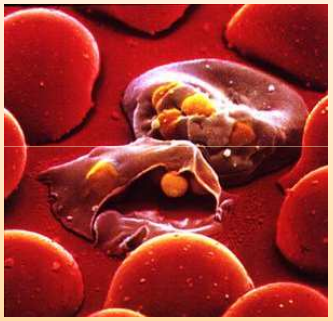
issus du premier hôte

issus de l'hôte final

issus de l'algue rouge

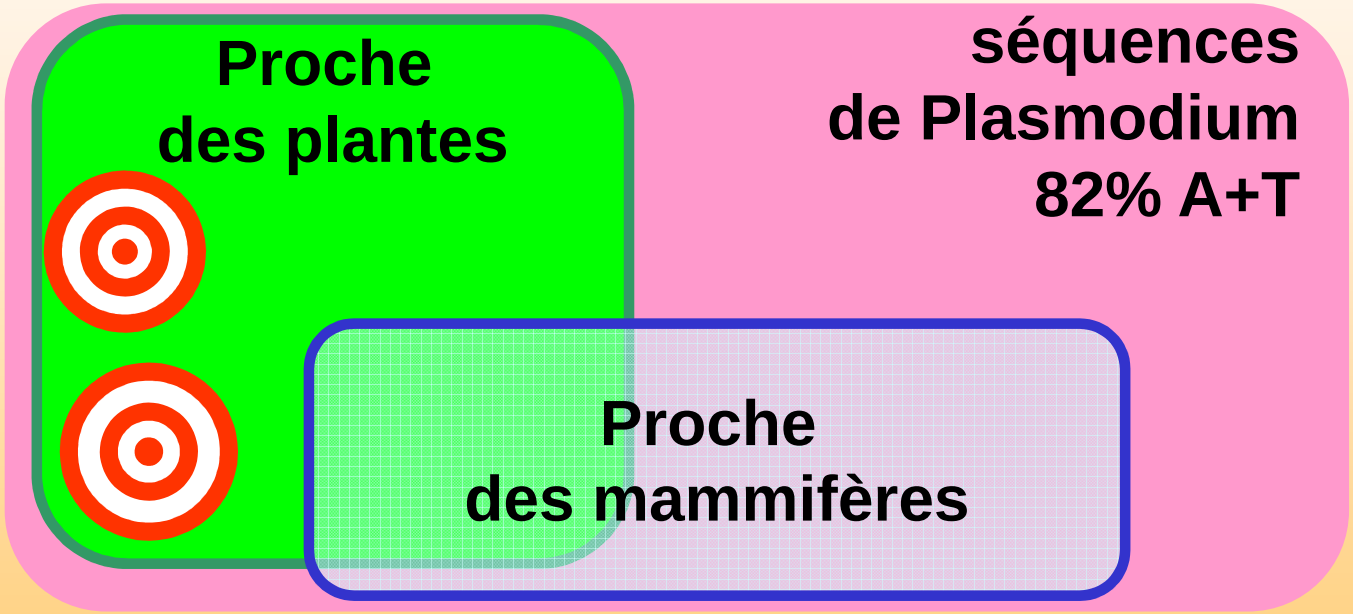
=> Sous-génome végétale ?

Plasmodium falciparum:
l ’agent du paludisme



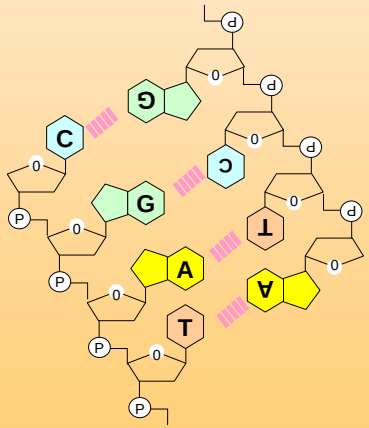
Médicaments herbicides

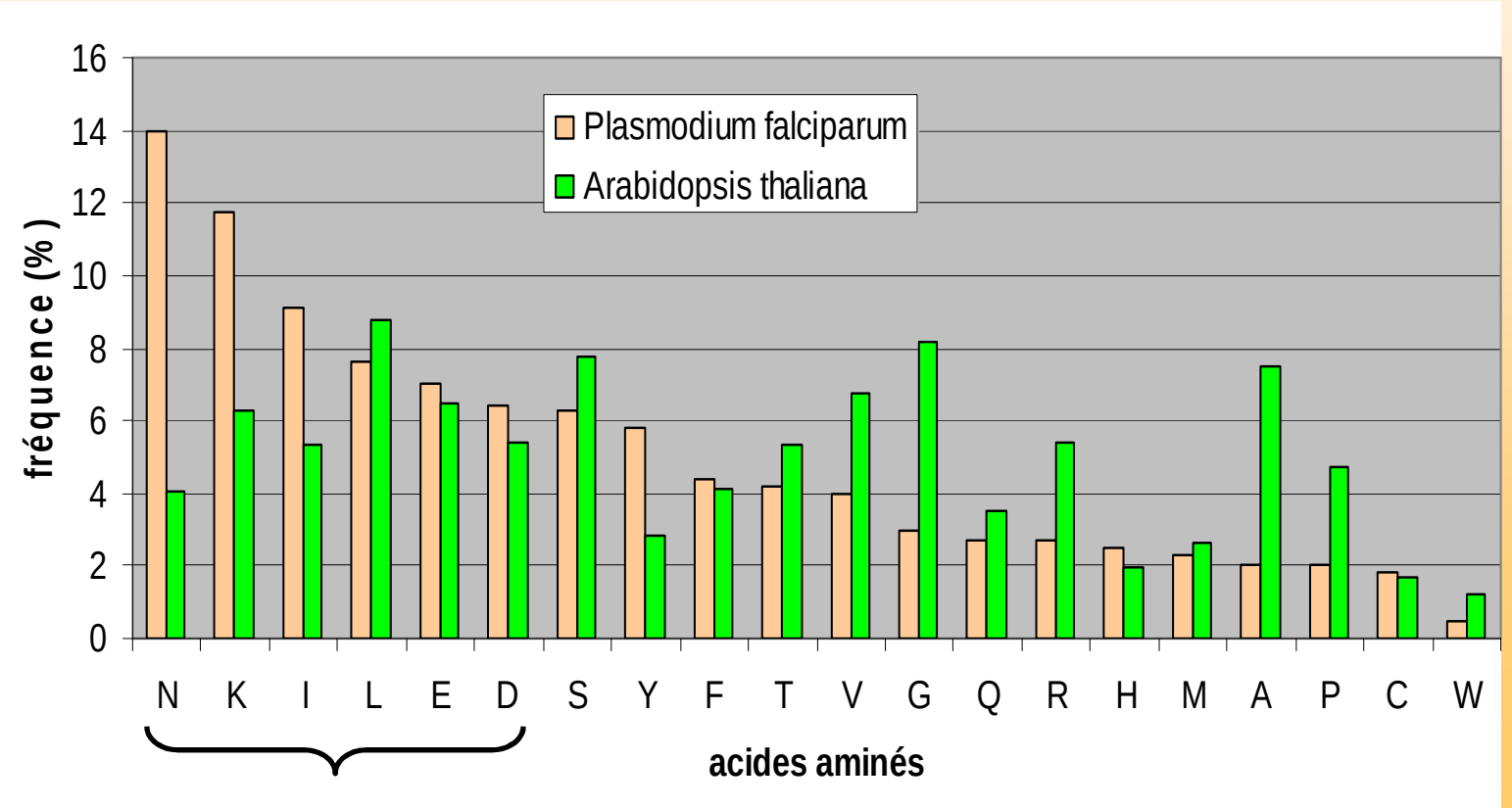
Analyse comparée des génomes d'*Arabidopsis thaliana*, de *Plasmodium falciparum* et de l'homme



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

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





Hypothèses

Le biais en aminoacides entraîne, si il n'est pas tenu en compte dans les modèles:

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 une conséquence du biais en acides nucléiques

---

4- La recherche du sous-génome végétal de *Plasmodium* (recherche de séquences homologues d'*Arabidopsis* proche du point de vue phylogénétique) doit tenir compte de **1** et **2**.

Plan

- Problématique générale

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence dans le cadre de la théorie de la fiabilité

- Conclusion générale

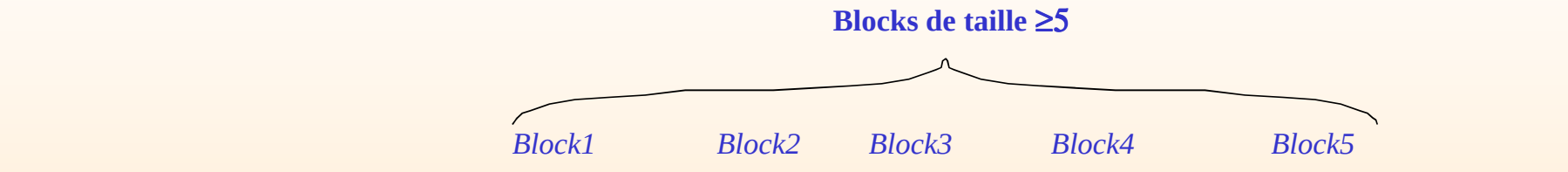
La parenté évolutive de séquences primaires est mesurée grâce à des alignements (1)

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)




```
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)

j dans la séquence **requête**
aligné avec ***k*** dans la séquence **sujette**,
avec une fréquence q_{jk}

$$S_{jk} = \lambda \log\left(\frac{q_{jk}}{p_j p_k}\right)$$

Principe de la mesure d'un alignement (3)

On a alors:

$$\begin{cases} q_{jk} \geq p_j p_k \Rightarrow S_{jk} \geq 0 & \text{Substitution favorable} \\ q_{jk} \leq p_j p_k \Rightarrow S_{jk} \leq 0 & \text{Substitution défavorable} \end{cases}$$

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 résidus

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 soient pas trop dissimilaires "
- Les séquences ont des tailles "comparables"

=> Hypothèses violées dans le cas général de la comparaison inter et intra génomes

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^* => $s(a,b^*)$

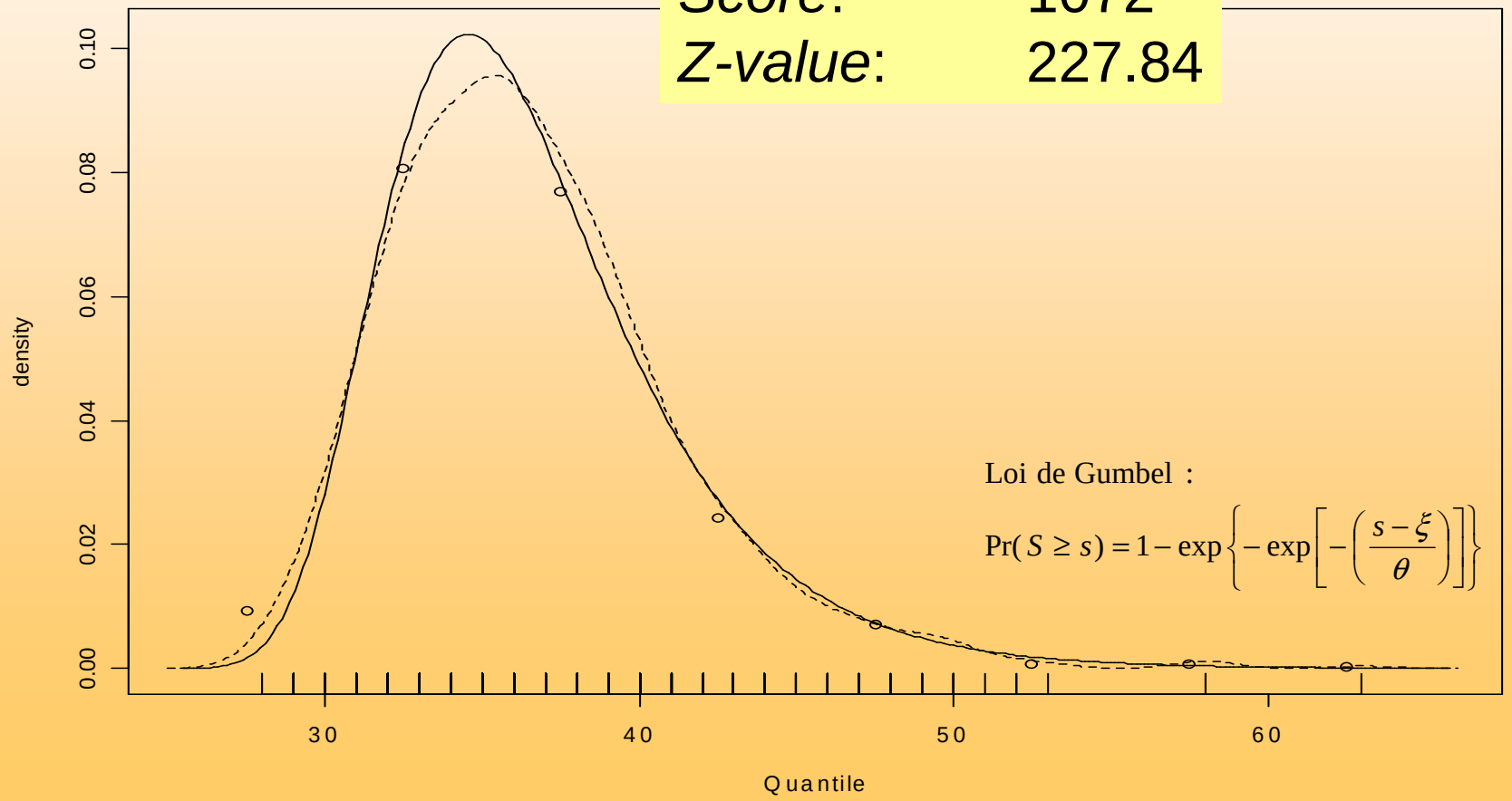
3- On observe la distribution des 1000 $s(a,b^*)$, ou se situe $s(a,b)$ dans cette distribution?

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

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



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 que 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:*

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

Signification théorique de la Z-value (2)
théorème TULIP

On considère une variable aléatoire $S(a,b^*)$ avec $\mu = E[S(a,b^*)]$ et σ^2 finies.

Quelque soit $k>1$ l'inégalité de Bienaymé-Chebyshev (Bienaymé, 1853; Chebyshev, 1867) énonce:

$$P\{|S(a,b^*)-\mu| \geq k\sigma\} \leq \frac{1}{k^2}$$

Considèrant uniquement la partie droite de la distribution, c'est à dire tel que $S(a,b^*) > \mu$, on a:

$$P\{S(a,b^*)-\mu \geq k\sigma\} \leq P\{|S(a,b^*)-\mu| \geq k\sigma\} \leq \frac{1}{k^2}$$

Et donc $$P\{S(a,b^*) \geq \mu + k\sigma\} \leq \frac{1}{k^2} \quad (1)$$

De plus:

$$s(a,b) \geq \mu + k\sigma \Rightarrow P\{S(a,b^*) \geq s(a,b)\} \leq P\{S(a,b^*) \geq \mu + k\sigma\} \quad (2)$$

Corrolaire 2 de TULIP

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

alors $z(a,b^*)$ (noté z) est la borne supérieure de $k \in]0,+\infty[$ tel que l'inégalité du T.U.L.I.P. (i.e., $P\big(S(a,b^*) \geq s(a,b)\big) \leq \frac{1}{k^2}$) soit vraie.

On a alors:

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

Plan

- Problématique générale

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence 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 $f(x,y): E \times E \to \Re^+$ 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 $f(x,y): E \times E \to \Re^+$ 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

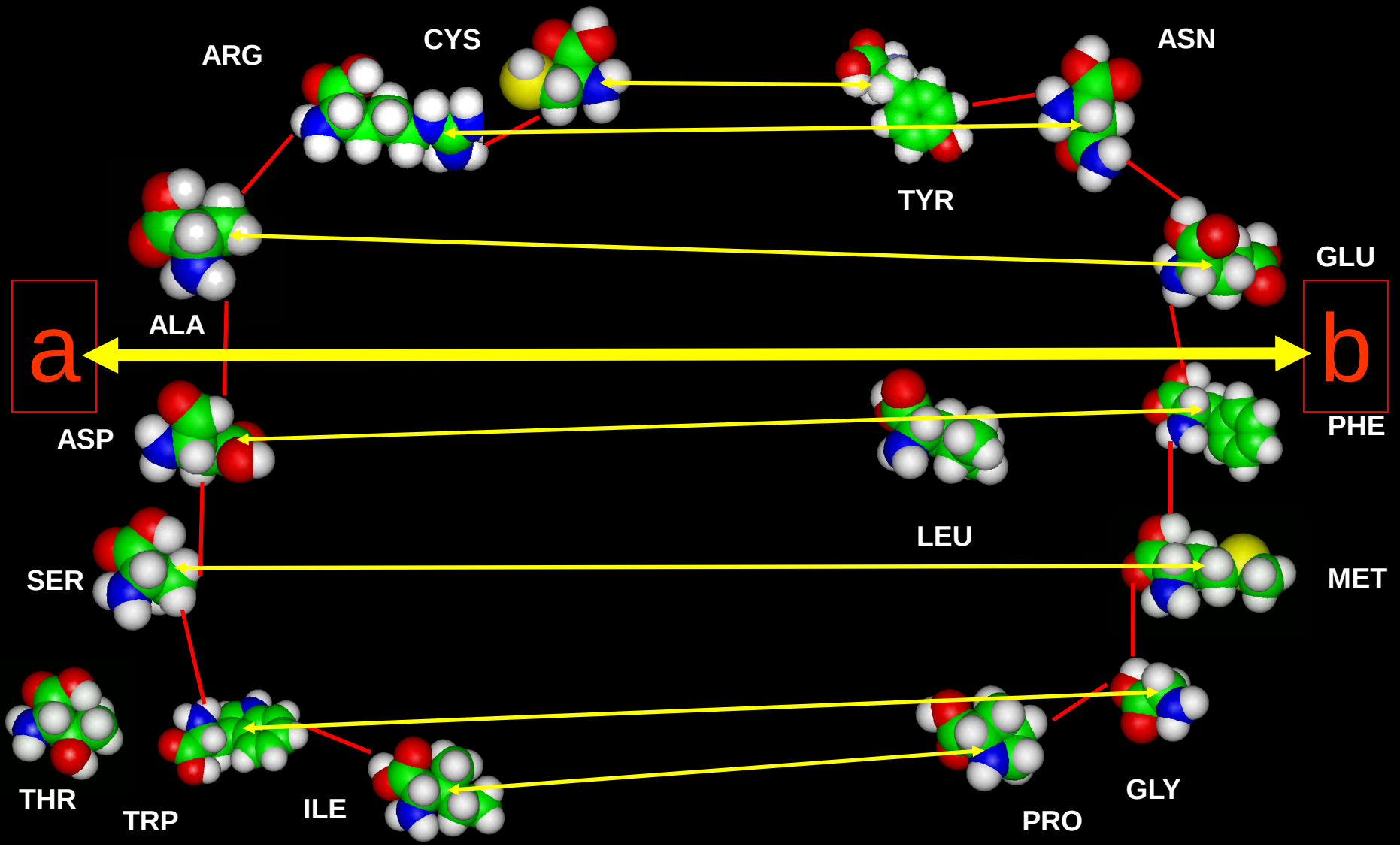
On appelle distance dans E toute fonction $d(x,y): E \times E \to \Re^+$ telle que

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)

$$s(j,k)=\log\frac{q_{jk}}{\pi_j\pi_k}$$

La théorie de l'information. Les bases.

Comment transmettre des données à moindre coûts (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.

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. Bien entendu cette surprise est difficilement chiffrable car elle varie d'un individu à l'autre. Les travaux de Claude Shannon (1948) ont permis de s'affranchir de cet aspect subjectif en liant l'information apportée par un événement E à sa probabilité de réalisation.

Cette théorie repose sur la définition de certaines quantités dont les deux plus importantes sont:

1- l'incertitude attaché à la possible réalisation d'un évènement E, notée h(E).

2- l'information mutuelle partagée par 2 évènements E et F, notée I(E;F).

Les matrices de substitution sont des matrices d'*informations mutuelles* (2)

- Soit un espace probabilisé $(\Omega, \Im, P)$

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

$$h(E) = -\log(P(E))$$, mesure l'information sur le système (ici Ω) apportée par l'occurrence de E

- On montre que $h(E \cap F) = h(E) + h(F)$ Si E et F sont indépendant

- ***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 $I_{F \to E} = I_{E \to F} \overset{def}{\equiv} I(E;F)$ *information mutuelle* entre E et F

- Avec les théorèmes classique de probabilités conditionnelles, on obtient

$$h(E \cap F) = h(E) + h(F) - I(E;F)$$

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

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

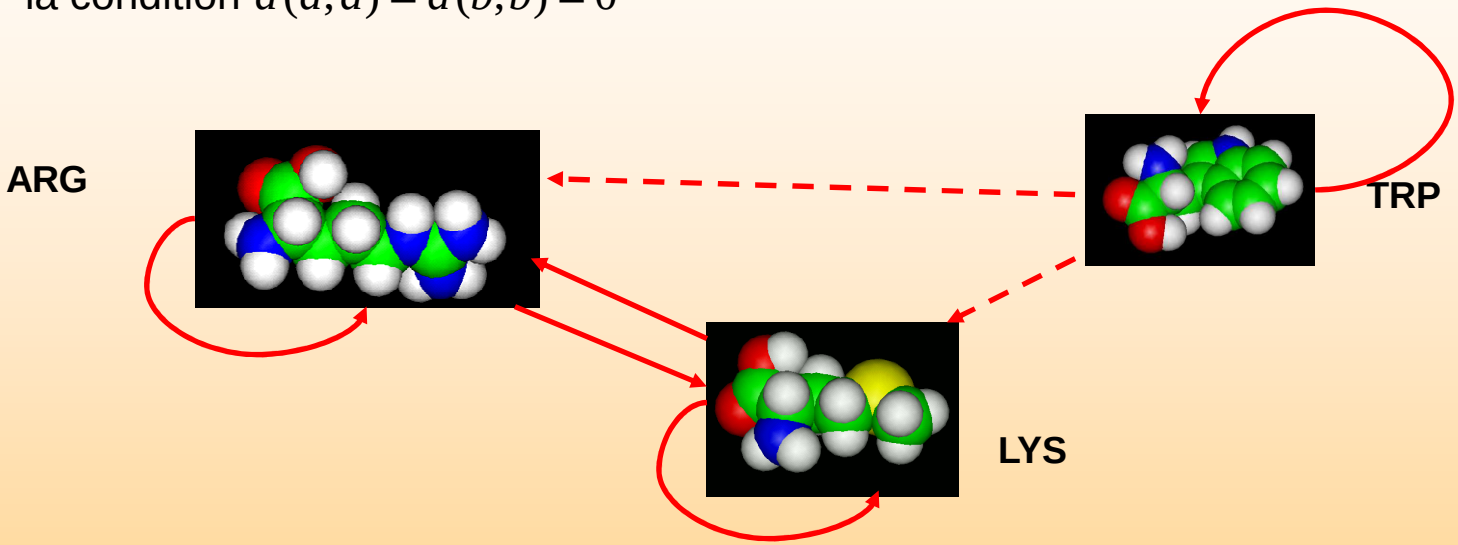
La façon dont on effectue la mesure « information mutuelle » sous-tend l'hypothèse d'indépendance des résidus

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 $d(a,a) = d(b,b) = 0$



- **Proposition** : Construire une distance entre acides aminés à partir d'une fonction composée

$$d(a,b) = \varphi \circ s(a,b)$$

conduit à une perte d'information mutuelle entre certains acides aminés. En conséquence, la distance entre acides aminés ne reflète pas entièrement leurs proximité en terme de propriétés structurales ou chimique. Cette proposition s'étend par définition aux distances entre séquences biologiques.

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

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

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

Soit Ω l'espace (i.e. un ensemble) des séquences biologiques, a et b deux séquences éléments de cet espace, alors le score de comparaison $s(a,b)$ est une similarité sur Ω.

On appelle q la quasi-dissimilarité associée à s. Ceci se fait grace à un théorème de passage.

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

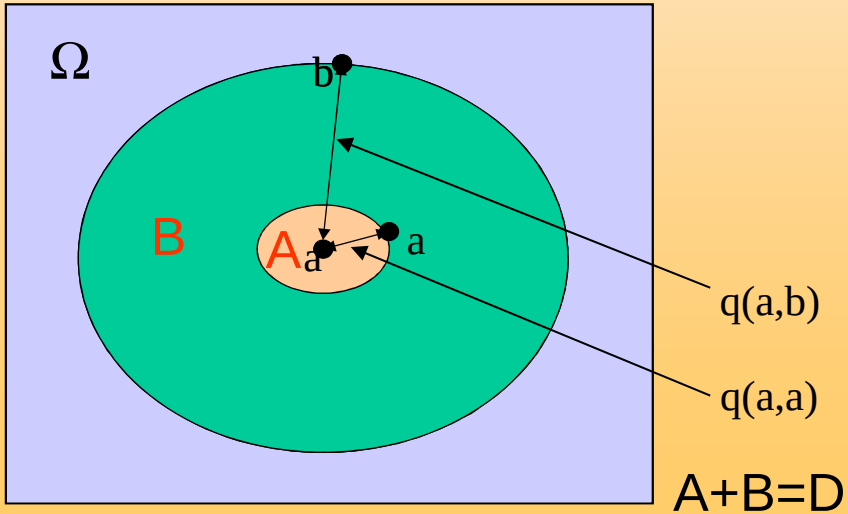
Corollaire

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

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

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

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



Plan

- Problématique générale

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence dans le cadre de la théorie de la fiabilité

- Conclusion générale

Modèle de la p-distance (1)

• 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 les mêmes résidus que a, connaissant la taille de a

• On munit le CSHP, l'espace des séquences d'une q-dissimilarité (celle associée à s). On définit deux variables aléatoires Q(a,b*)=exp(-S(a,b*)) et Q(b,a*)=exp(-1/S(b,a*))

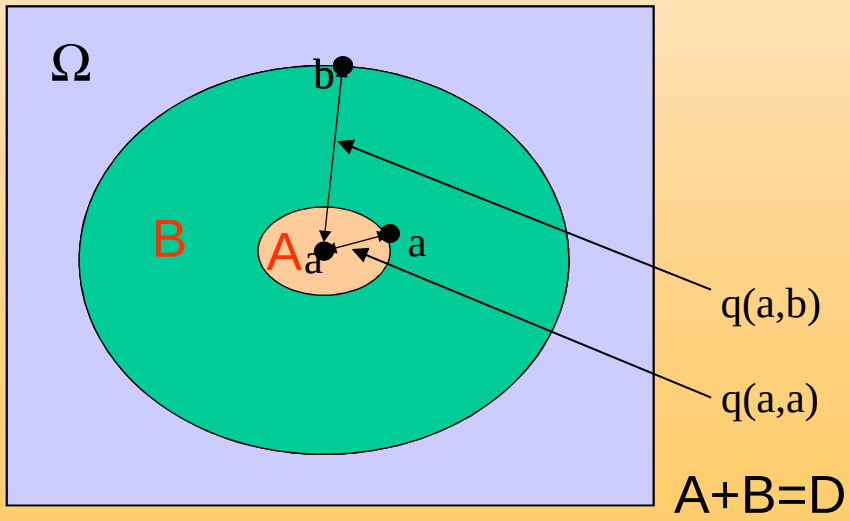
• P(Q(a,b*)<=x) est donc la probabilité pour que b* soit proche de a au plus de x, donc que b* partage certains résidus (ou certaines caractéristiques de ces résidus) avec a

Une nouvelle approche probabiliste (2)

- On peut alors définir:

$$P\{Q(a,b^*) \le q(a,a) / Q(a,b^*) \le q(a,b)\}$$

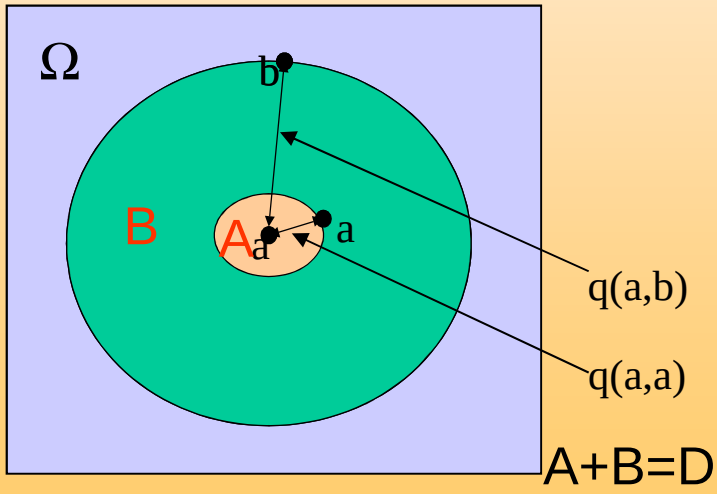
Probabilité pour que b* soit aussi proche de a que a l'est de lui-même, sachant que b* est au plus éloignée de a de q(a,b)



$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)}{P(D)}$$

Une nouvelle approche probabiliste (3)

$$P(A/D)=\frac{P\{Q(a,b^*)\leq q(a,a)\}}{P\{Q(a,b^*)\leq q(a,b)\}}\leq\frac{z^2(a,b^*)}{z^2(a,a^*)}$$



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 (4)

Prise en compte des deux origines:

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

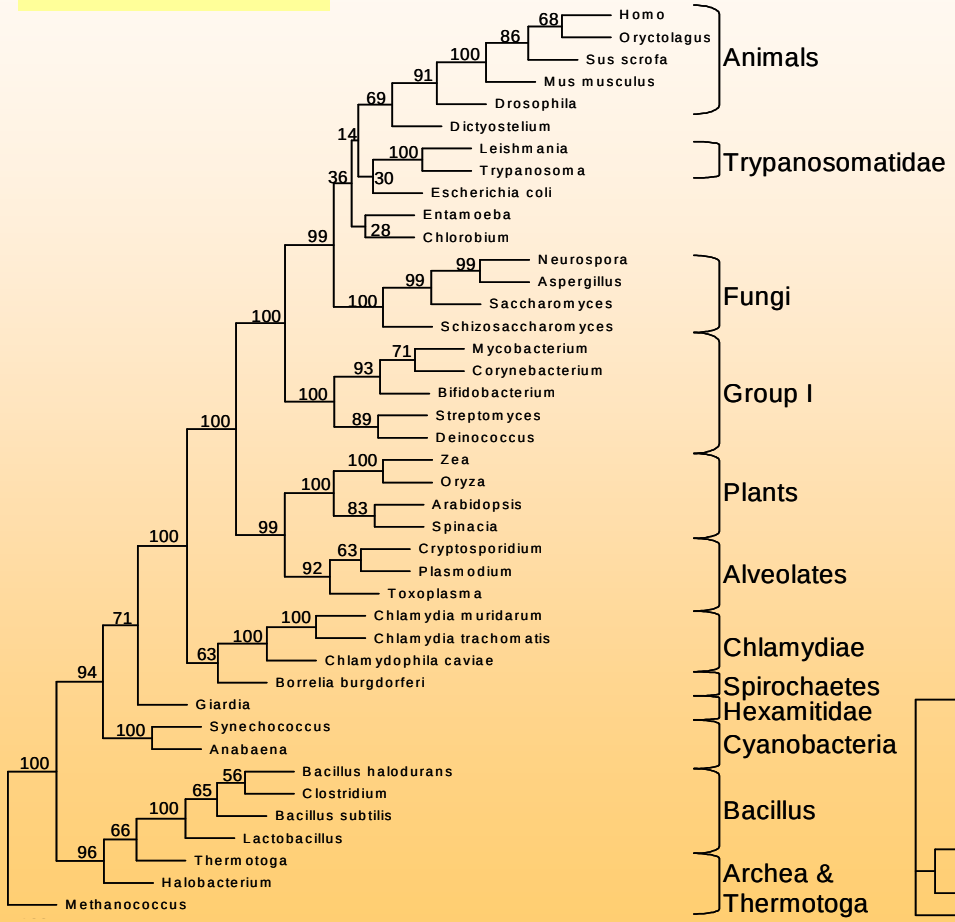
$$t_{\beta,b} = -\log\left(\frac{z^2(b^*,a)}{z^2(b,b)}\right)$$

Calcul final de la distance évolutive:

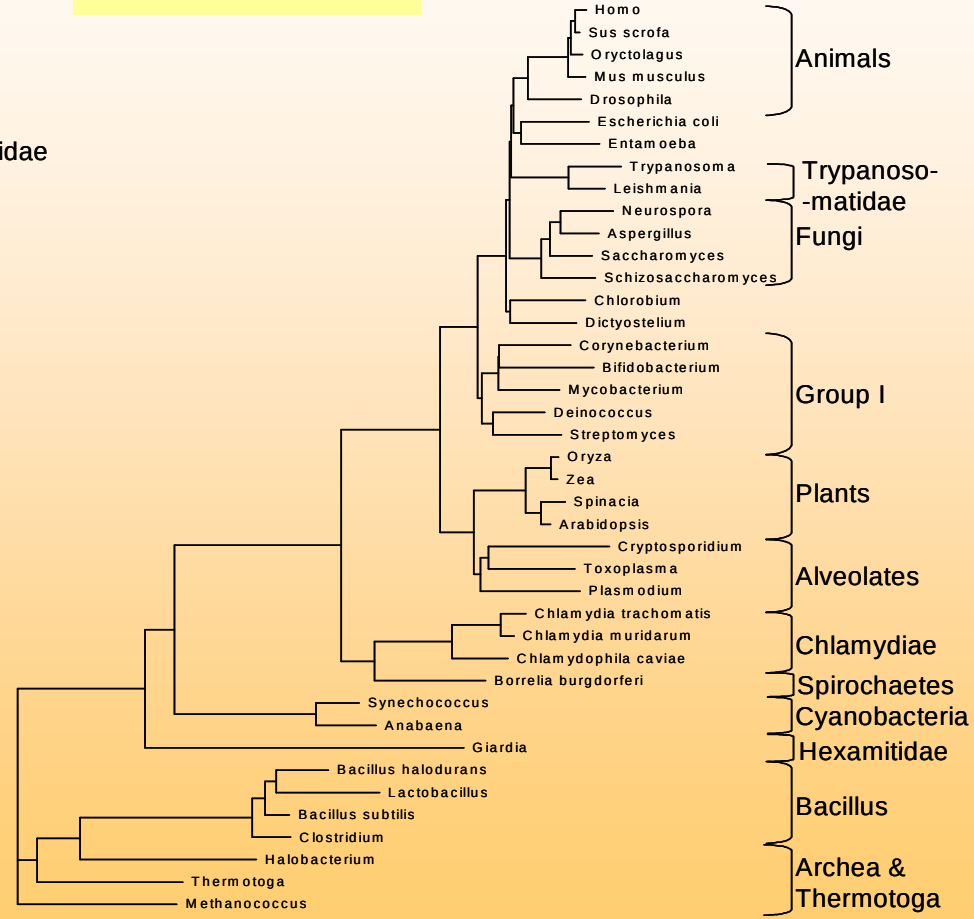
$$t_\beta = \left(t_{\beta,a}^2 + t_{\beta,b}^2\right)^{1/2}$$

Exemple 1: Glucose-6-Phosphate Isomerase

MAB TREE

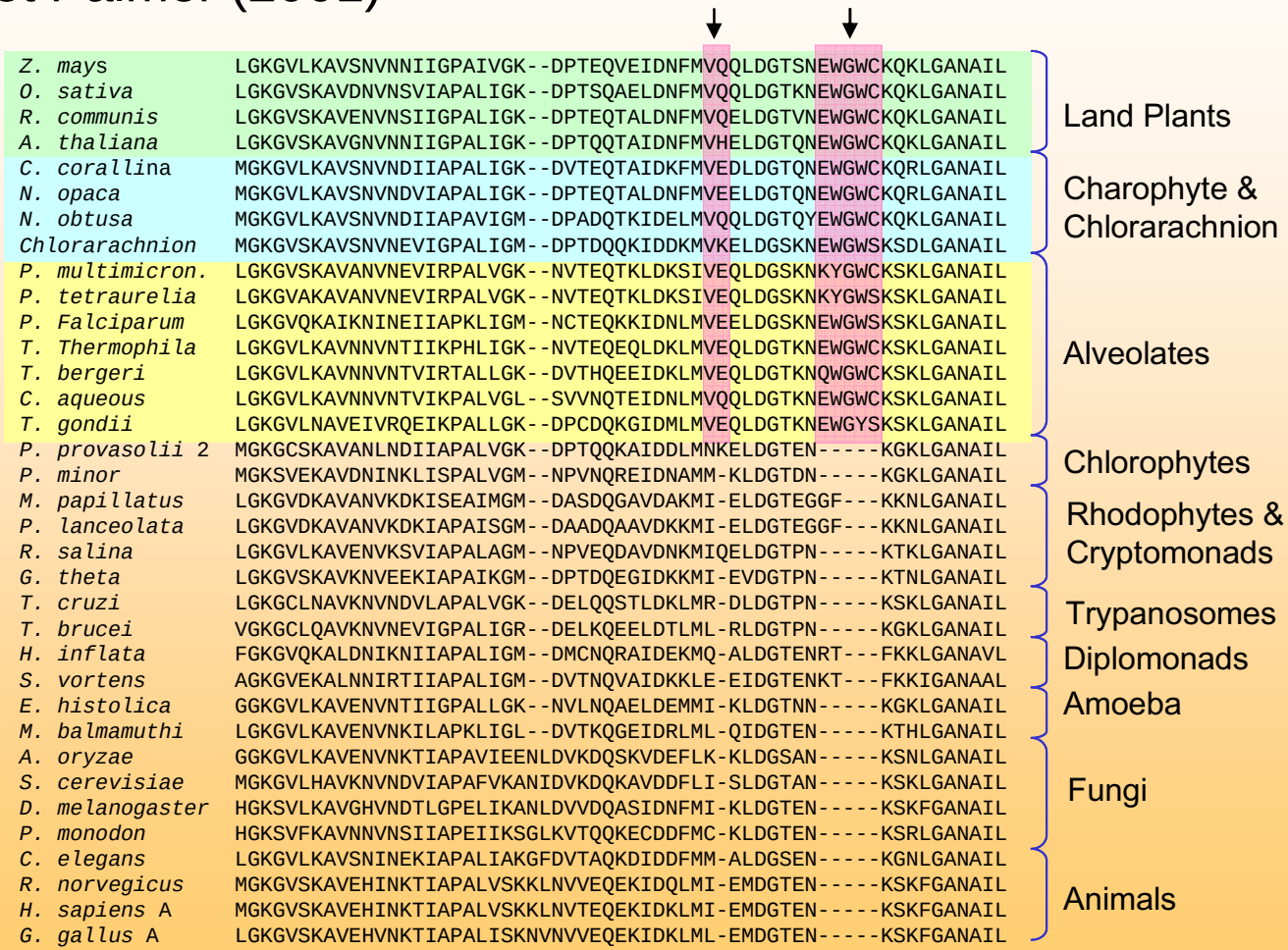


TULIP TREE



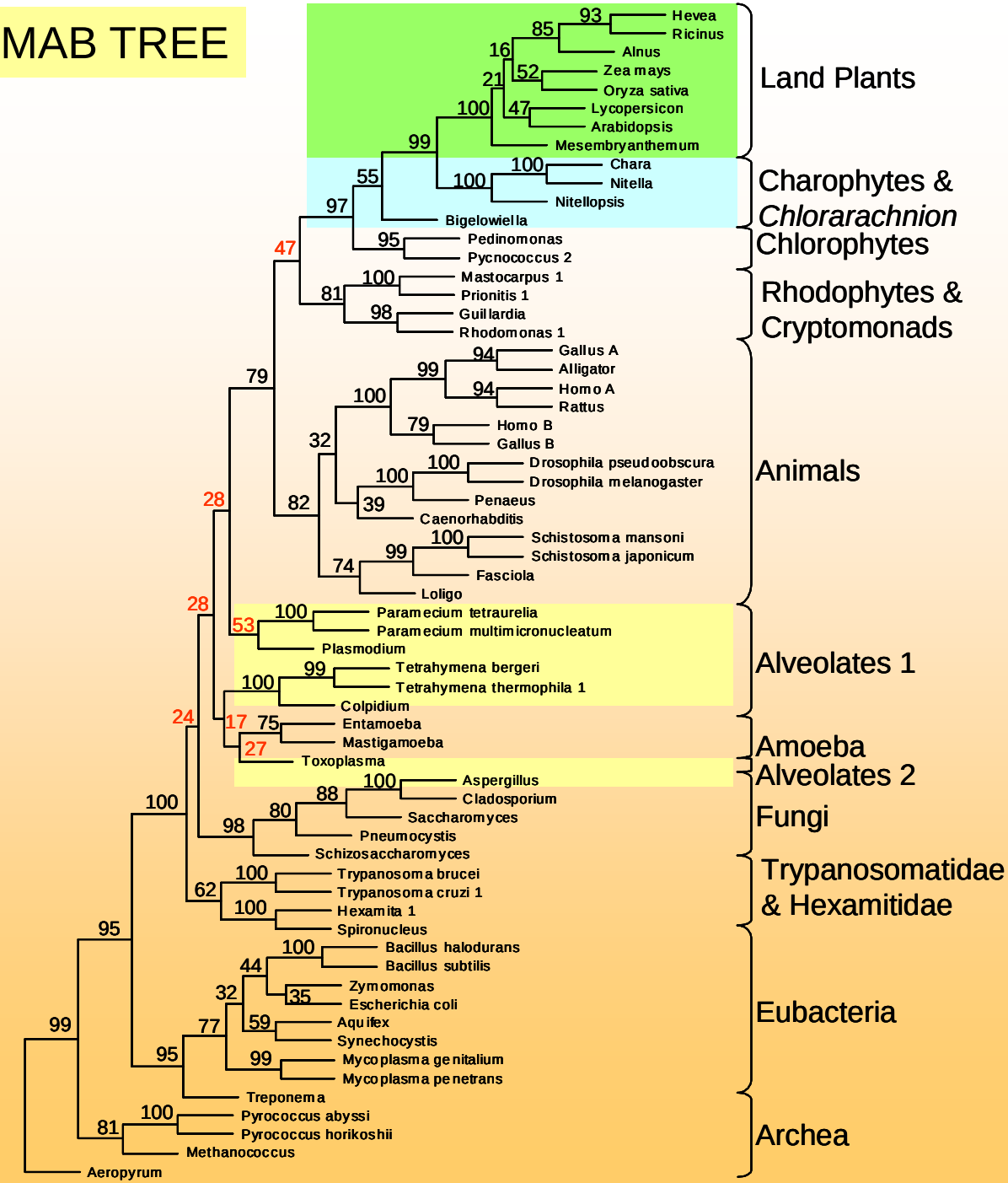
Exemple 2: l'énolase(1)

Keeling et Palmer (2001)



MAB TREE

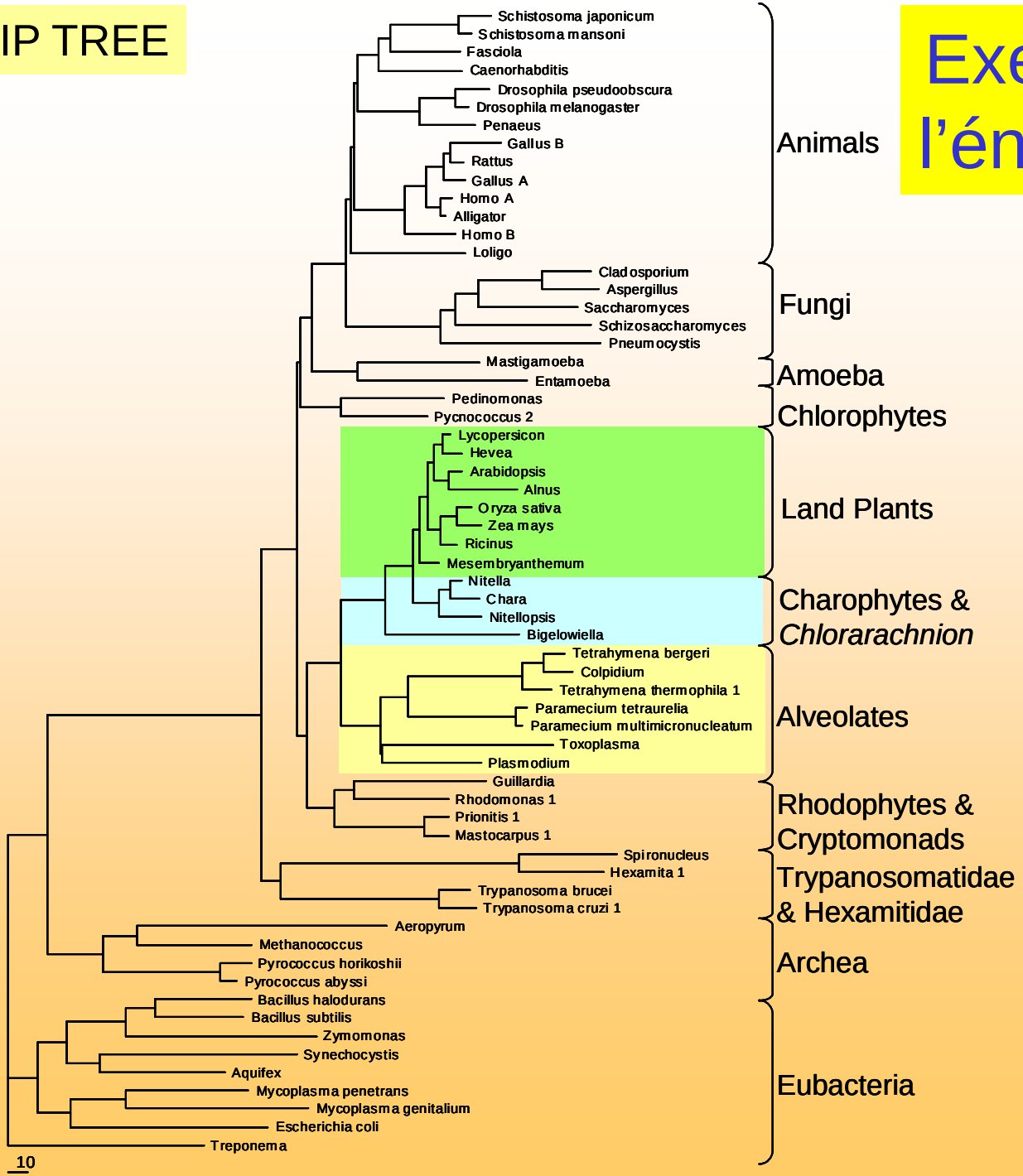
Exemple 2: l'énolase(2)



Bastien O., Ortet P., Roy S., and Maréchal E. (2005) BMC Bioinformatics 6:49

TULIP TREE

Exemple 2: l'énolase(3)



Plan

- Problématique générale

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence 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)

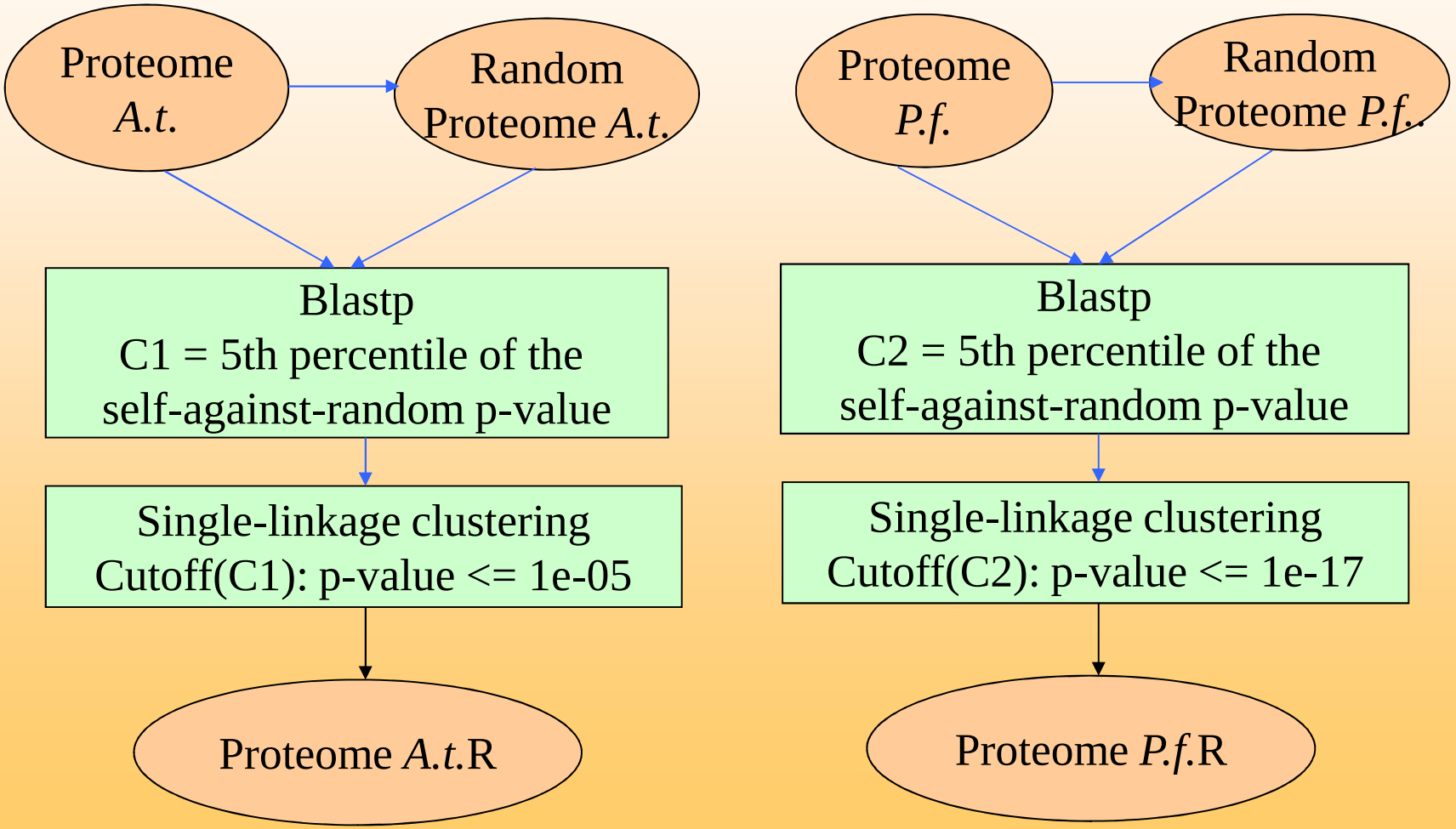
Idée générale

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

Bastien O., Lespinats S., Roy S., Métayer K., Fertil B., Codani J.J. and Maréchal E. (2004) Gene 336: 163-173

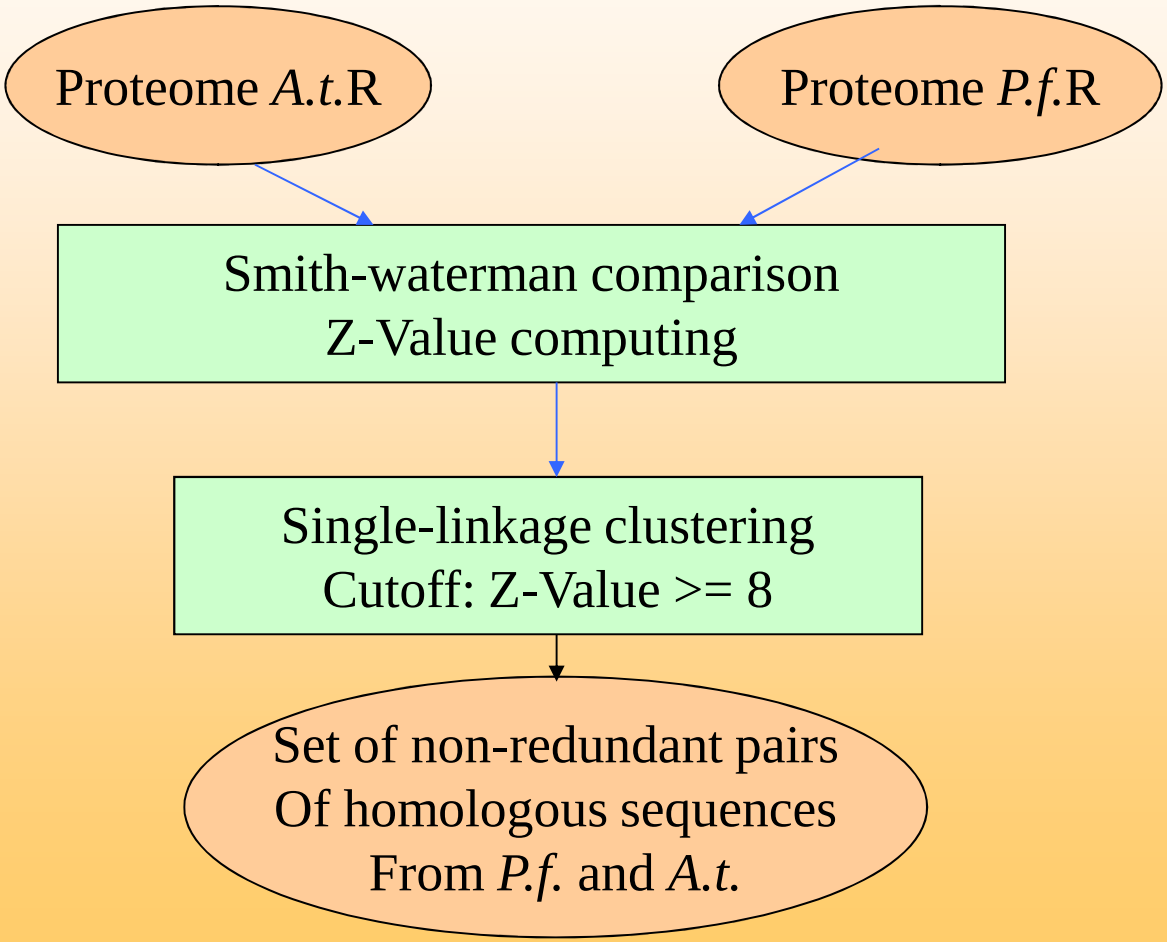
Comparaison des protéomes 3

étape 1: détermination des protéomes représentatifs



Comparaison des protéomes 4

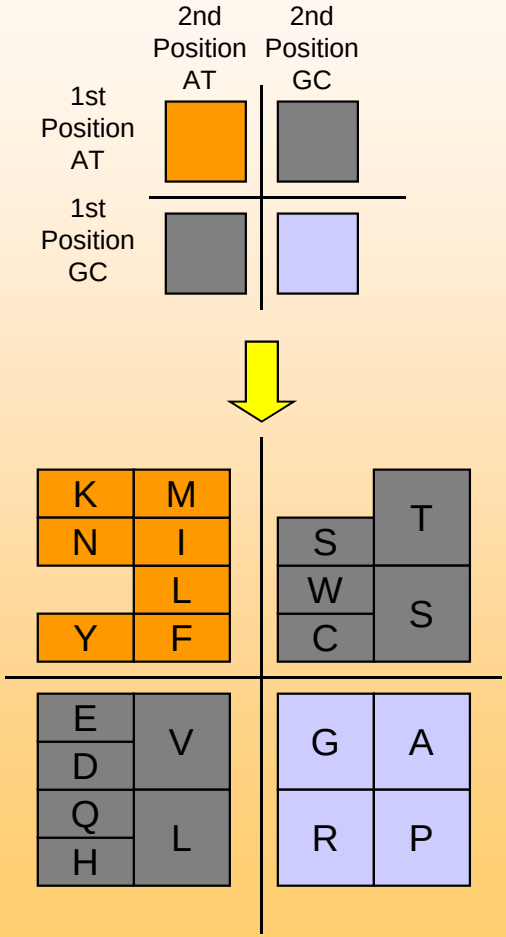
étape 2: comparaison des protéomes R



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

Partitionnement de la table des codons

Foster et al, 1997



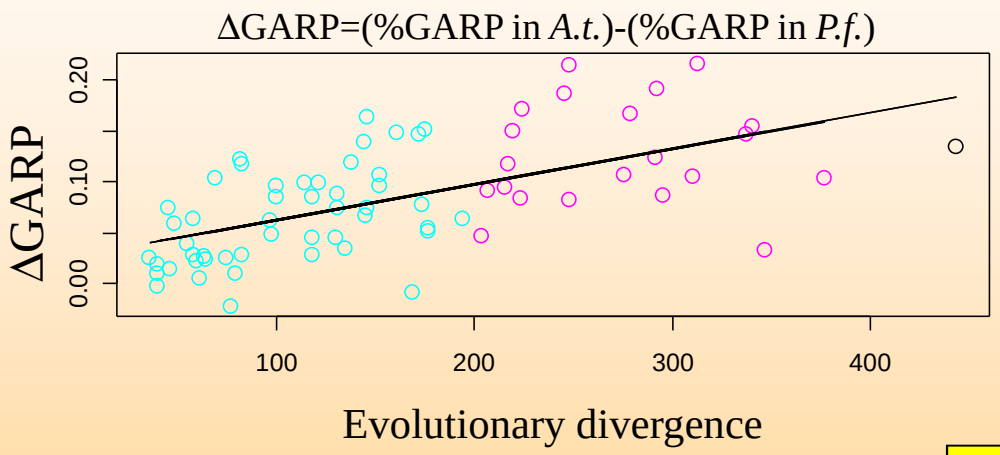
Hypothèses

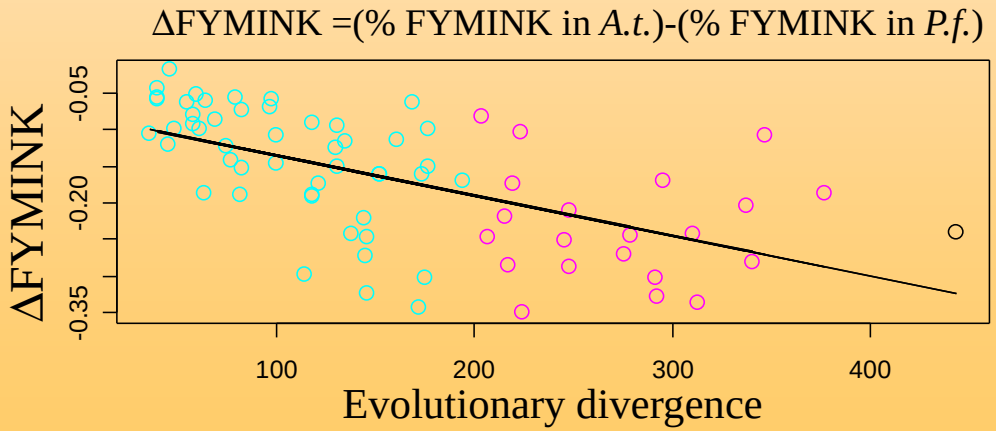
1- Les séquences riches en A+T vont coder pour des protéines enrichies en FYMINK par rapport à leurs homologues (moins riches en A+T).

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 *A.t./P.f.*

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



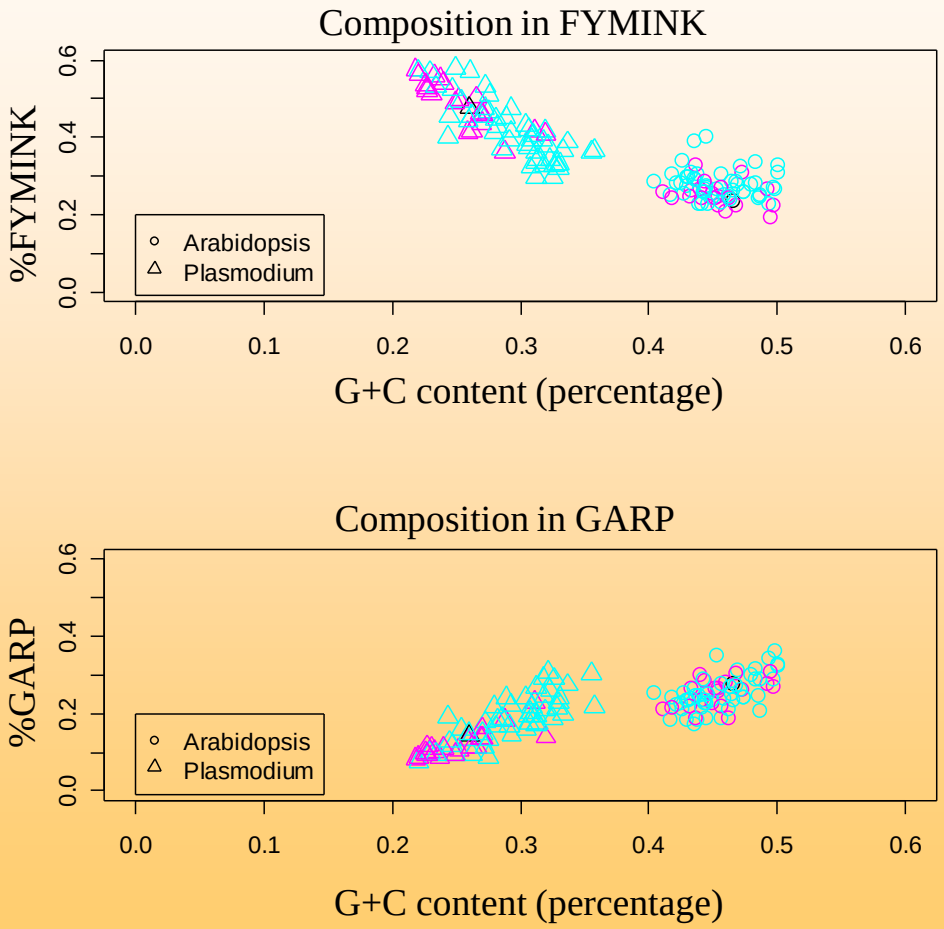


-Les protéomes de *A.t.* et *P.f.* ont des évolutions directionelles différentes.

- 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.*

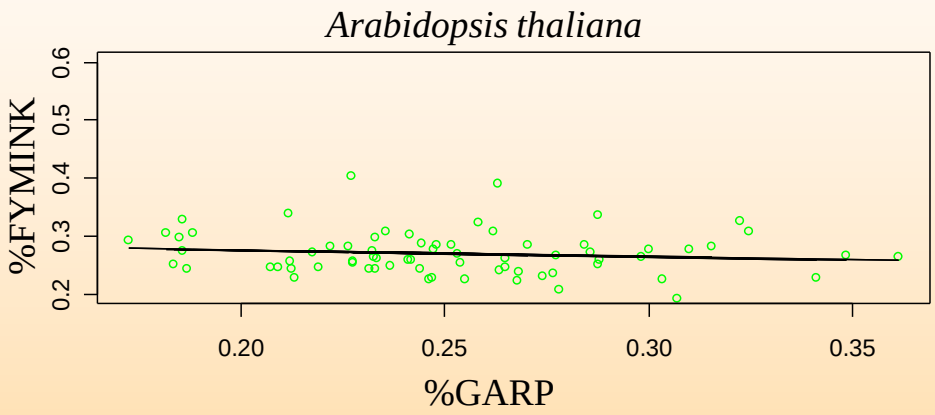


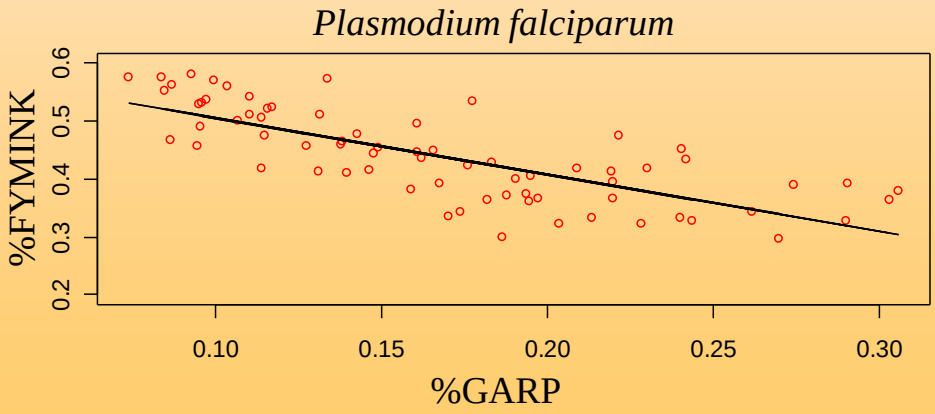
-Pas de recouvrement entre les compositions en G+C et en %GARP (ou %FYMINK) entre les 71 couples.

- 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}

$$dirAtPf_{jk} = \lambda . \ln\left(\frac{q_{j,k}}{\pi_j \tau_k}\right)$$

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

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

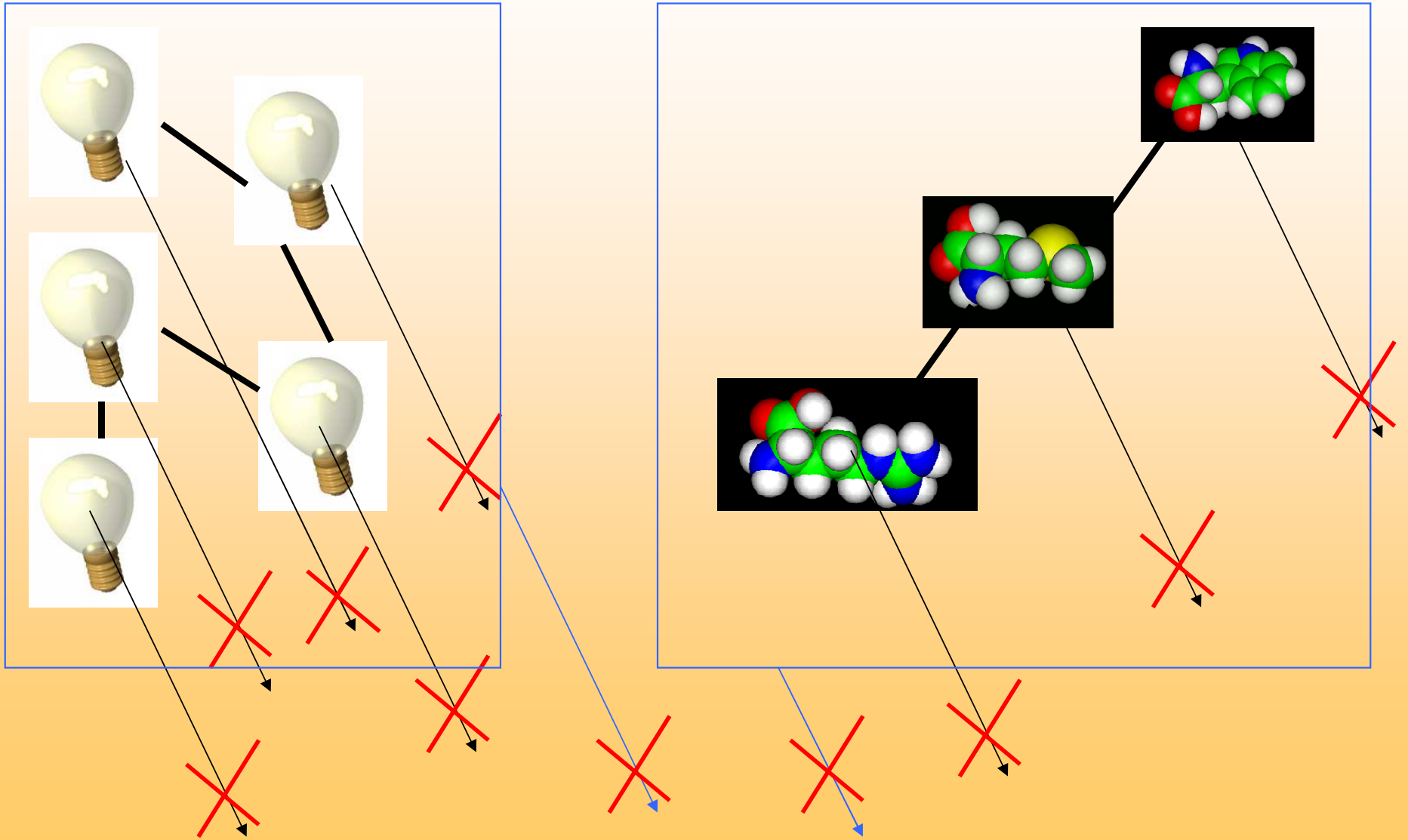
- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence dans le cadre de la théorie de la fiabilité

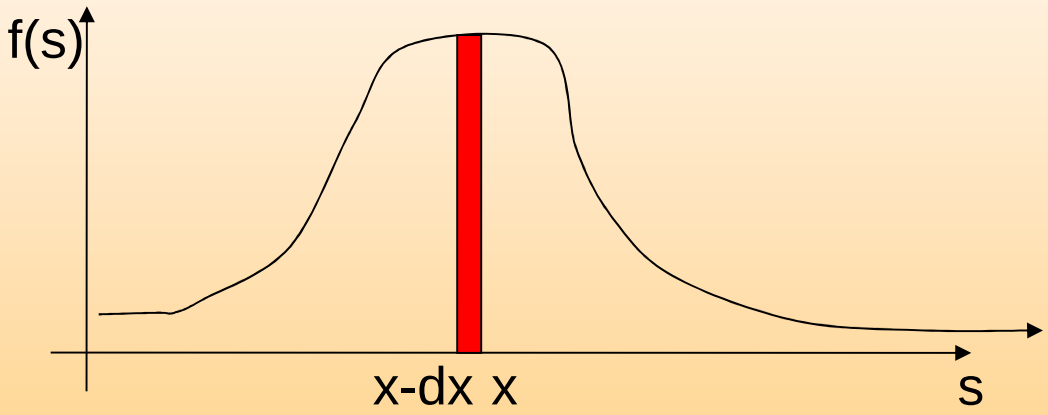
- Conclusion générale

La théorie de la fiabilité



La fonction de longévité (1)

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



On considère la probabilité d'être proche de x par valeur inférieure par unité de x.

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

La fonction de longévité (2)

Théorème: La fonction de longévité est le rapport de la densité de probabilité sur la fonction de distribution. C'est à dire:

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

Corollaire: La fonction de longévité, la fonction de densité et la fonction de distribution sont trois concepts équivalents.

$$P(X\leq x)=\exp(-\int_{x}^{+\infty}\psi(u)du$$

L’évolution des monomères en fonction du temps (1)

- L’information mutuelle entre deux séquences , fonction additive, représente la magnitude de la redondance d’information entre les séquences au niveau monomérique.

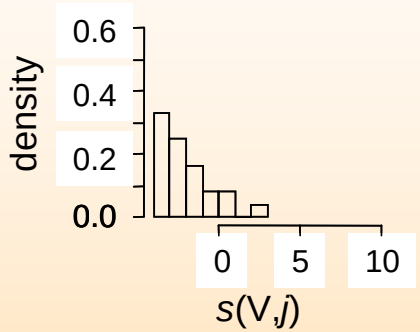
- Statistiquement, l’information mutuelle entre deux acides aminés décroît avec le temps

- On observe que

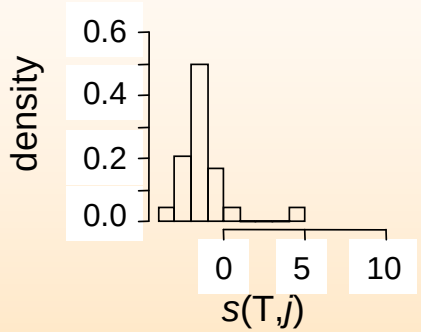
$$P_i(S_i \leq s_i) = 1 - \exp(-\lambda . s_i)$$

L'évolution des monomères en fonction du temps (2)

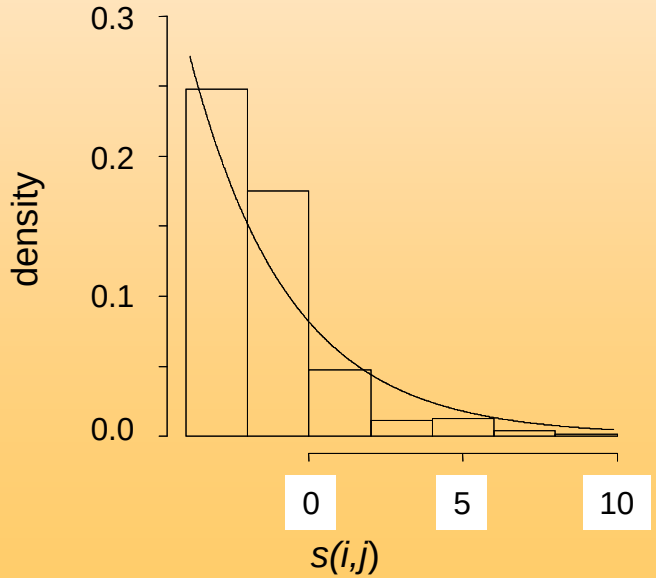
A. Valine: a non-aging component



B. Threonine: an aging component



C. All residues (based on BLOSUM 62): non-aging components



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

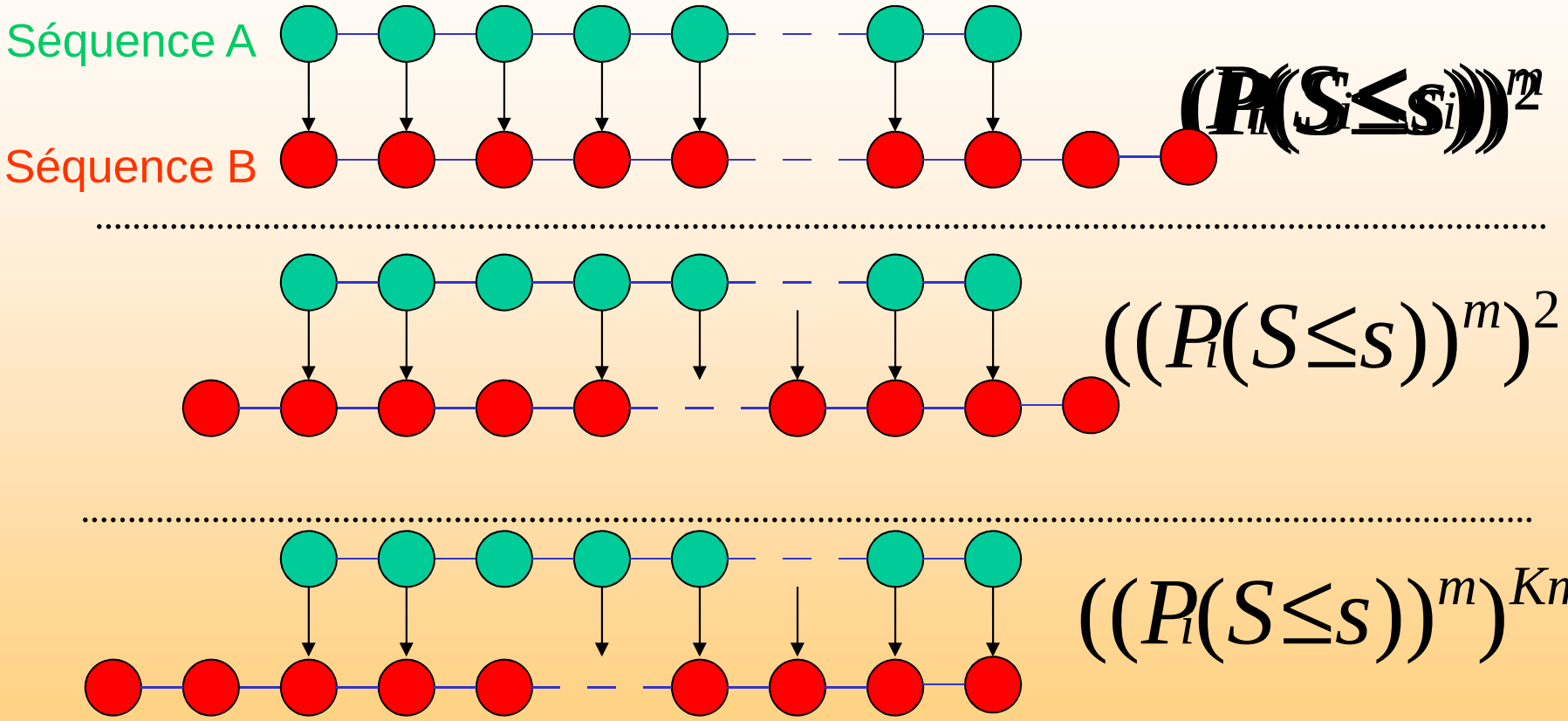
- On défini les approximations suivante, ou m est la taille de l'alignement entre a et b

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

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

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

Le calcul de la loi de probabilité (1)



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

Le calcul de la loi de probabilité (2)

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

On déduit celle de la densité de probabilité, $f(s) = \frac{\partial F}{\partial x}(s)$

Ce qui permet le calcul de la fonction de longévité $\psi(x) = \frac{f(x)}{F(x)}$

Et donc: $$\psi(x) = \frac{f(x)}{F(x)} = \frac{Kmnf_i(x)}{(P_i(S \leq s))}$$

Le calcul de la loi de probabilité (3)

On remplace dans l'expression précédente la forme de la loi de la probabilité de l'évolution des résidus, $P(S \leq s) = 1 - \exp(-\lambda . s)$

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

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

En utilisant la formule $P(X \leq x) = \exp(-\int_{x}^{+\infty} \psi(u)du$, on obtient la distribution de probabilité:

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

Où les paramètres ont une signification explicite

La loi de probabilité du z-score est indépendante de la composition et de la taille des séquences(3)

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 du z-score:

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

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

- Le Z-Score comme estimation de la significativité d'un score d'alignement dans le cas de la comparaison de deux séquences quelconques

- La comparaison de séquence dans le cadre de la théorie de l'information

- Le CSHP comme modèle général permettant le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques

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

- La comparaison de séquence dans le cadre de la théorie de la fiabilité

- Conclusion générale

Conclusion générale

- La théorie de l'information fourni un cadre adapté pour reformuler certains principes néo-Darwinien dans des termes mathématiques. L'alignement de séquence replacer 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 voulut au départ du projet de thèse

- La théorie du z-score permet soit un accès à une majoration, soit un calcul plus précis de la probabilité « d'alignement ». Comme toute méthode Monté-Carlo, elle coûte cher en temps de calcul et un modèle de z-Blast est en développement.

Et maintenant, l'exploration des génomes!!!!

Remerciements

Laboratoire de Physiologie Cellulaire Végétale (CEA Grenoble)
Eric Maréchal
Mais aussi Juliette, Maryse, Cyrille, Jacques, Jess, Pascaline, Gilles, Samuel, Stéphane M., Stéphane R., Alex, Florent et tous les autres membres du laboratoire!!

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

Gene-It
Jean-Jacques Codani
Karine Metayer

Laboratoire Imagerie Médicale quantitative
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
Cordélia Bisanz
Marie-France Cesbron

Et aussi mes amis, ma famille et Delphine