Le théorème TULIP et l'espace de configuration des protéines homologues

Olivier Bastien

UMR 5168 Grenoble

Plan

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

- L'**espace de configuration des protéines homologues (CSHP)** comme modèle permettant le calcul de distance évolutive entre séquences et la reconstruction **d'arbres phylogénétiques**.

- Conclusion générale

Partie I

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

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

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

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)

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

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

globale

A chaque résidus

L'alignement optimal est celui qui maximise le score

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

1- Classiquement: estimation de la probabilité d'obtenir un scor
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 séquence 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

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

Évaluation de la pertinence d'un score: Le Z-Score (1)

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?

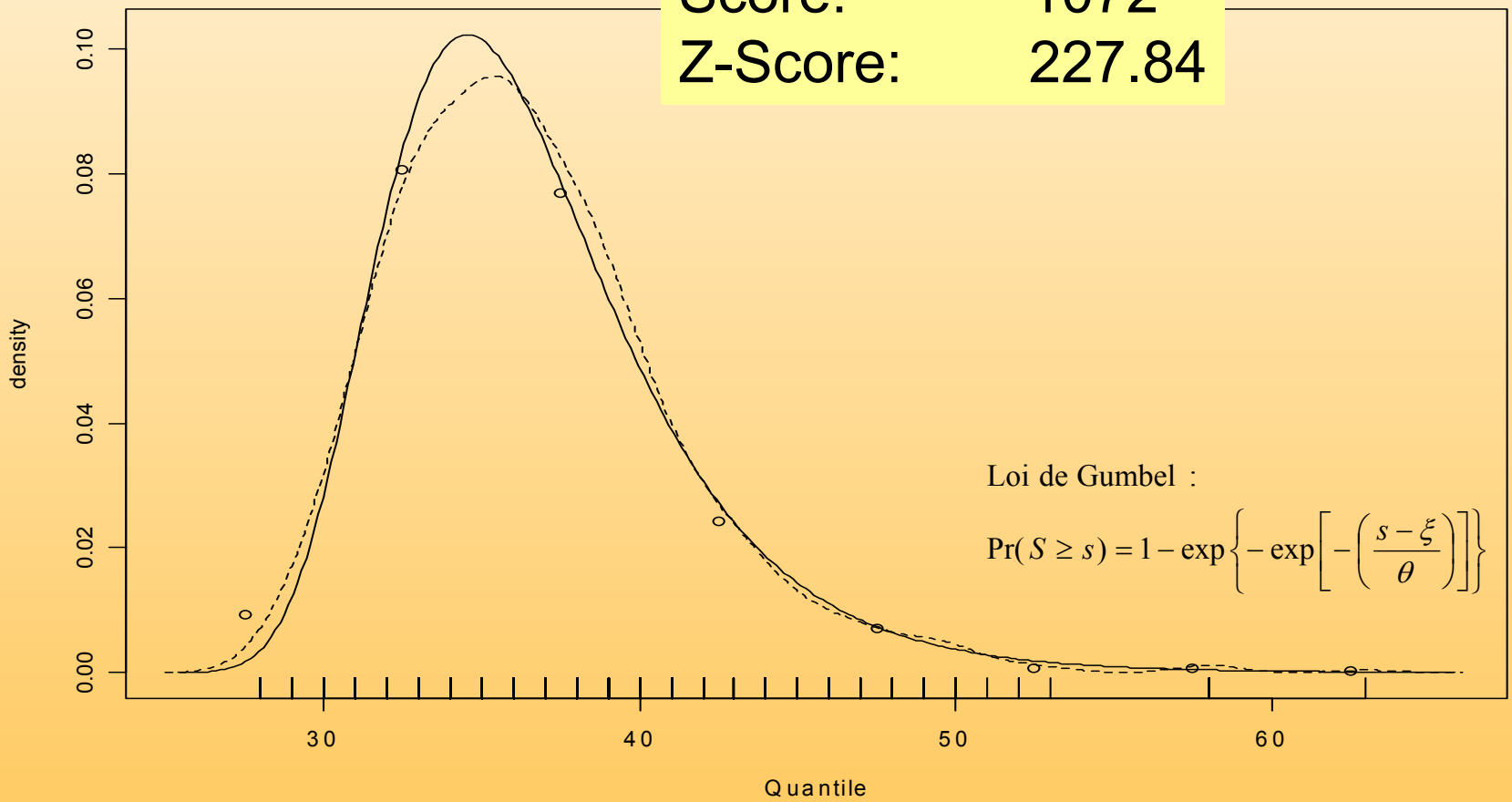
$$Z - score = \frac{s(a,b) - E[S(a,b^*)]}{\sigma}$$

4) 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.

Evaluation de la pertinence d'un score: Le Z-Score (2)

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

Score: 1072
Z-Score: 227.84



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*)>=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:*

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

Bastien O., Aude J.C., Roy S. and Maréchal E. (2004) Bioinformatics 20: 534-537

Signification théorique du Z-Score (2)
Corollaire 2 de TULIP

Soit $z(a,b^*) = \frac{s(a,b) - E[S(a,b^*)]}{\sigma[S(a,b^*)]}$, 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., $P(S(a,b^*) \geq s(a,b)) \leq \frac{1}{k^2}$) soit vraie ; On a alors

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

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

Rappel sur les notions de proximité

Similarité

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

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

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

Dissimilarité

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

$$\textbf{i)} \quad \forall x \in E, \forall y \in E, f(x,y) = 0 \Leftrightarrow x = y$$

$$\textbf{ii)} \quad \forall x \in E, \forall y \in E, f(x,y) = f(y,x)$$

Distance

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

$$\textbf{i)} \quad \forall x \in E, \forall y \in E, d(x,y) = 0 \Leftrightarrow x = y$$

$$\textbf{ii)} \quad \forall x \in E, \forall y \in E, d(x,y) = d(y,x)$$

$$\textbf{iii)} \quad \forall x \in E, \forall y \in E, \forall z \in E, d(x,z) \le d(x,y) + d(y,z)$$

L'espace des acides aminés

-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 acides aminés 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.

- 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**: 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 $I_{F \to E} = I_{E \to F}$ et l'information mutuelle entre E et F est notée $I(E;F)$

- 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

- En décomposant $s(i,j) = \log \frac{q_{ij}}{\pi_i \pi_j}$, il vient

$$s(i,j) = \log(P_{al}(i \cap j)) - \log(P_{\pi}(i)) - \log(P_{\pi}(j))$$

et donc

$$s(i,j) = h(i) + h(j) - h(i \cap j)$$

Ce qui amène

$$s(i,j) = I(i;j)$$

Conclusion qui s'étend immédiatement au score entre séquence

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

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

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

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

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

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 aref, la séquence référence.

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

- 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

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

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 :

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

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

Théorème Q (passage d'une similarité à une quasi-dissimilarité)

Soit f une similarité sur E, alors $q = e^{-f}$ est une quasi-dissimilarité. On dira que q est associée à f .

La démonstration est évidente sur les définitions.

Soit Ω l'espace (i.e. un ensemble) des séquences biologiques, x et y deux séquences éléments de cet espace, alors le score de comparaison $s(x,y)$ est une similarité sur Ω.
On appelle q la quasi-dissimilarité associée à s.

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

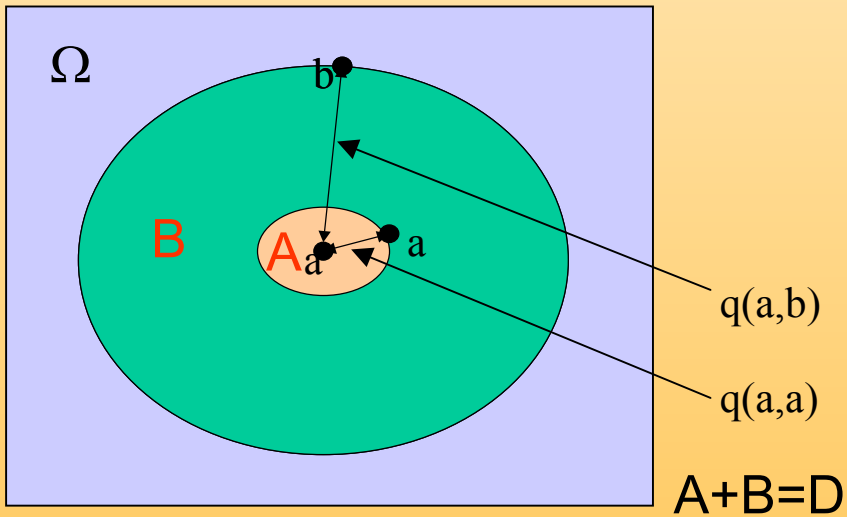
Corollaire

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

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

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

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



Partie III
Le CSHP permet le calcul de distance évolutive entre séquences et la reconstruction d'arbres phylogénétiques.

Modèle de la p-distance

Feng et al, 1985
Doolittle et al, 1996
Feng et Doolittle, 1997
Broccheri, 2001

• 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(-S(b,a*))

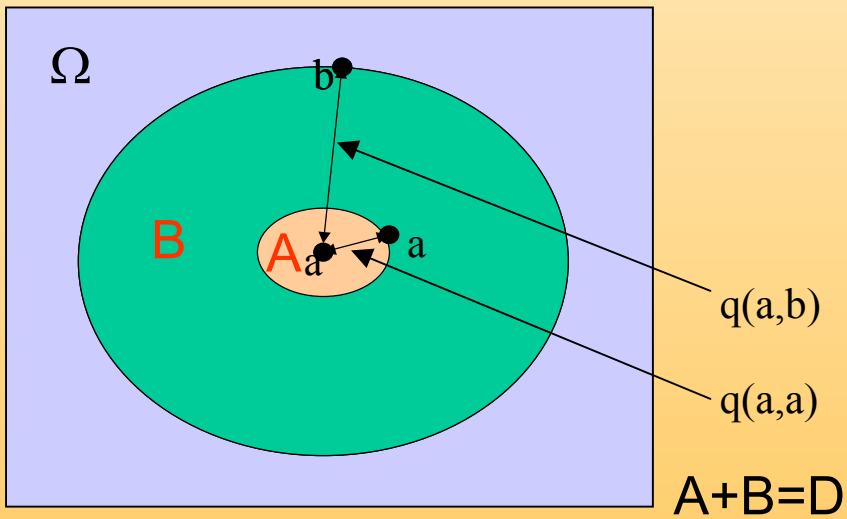
- P(Q(a,b*)<=x) est donc la probabilité pour que b* soit proche de a au plus de x, donc que b* partage certains résidus (ou certaines caractéristiques de ces résidus) avec a

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)

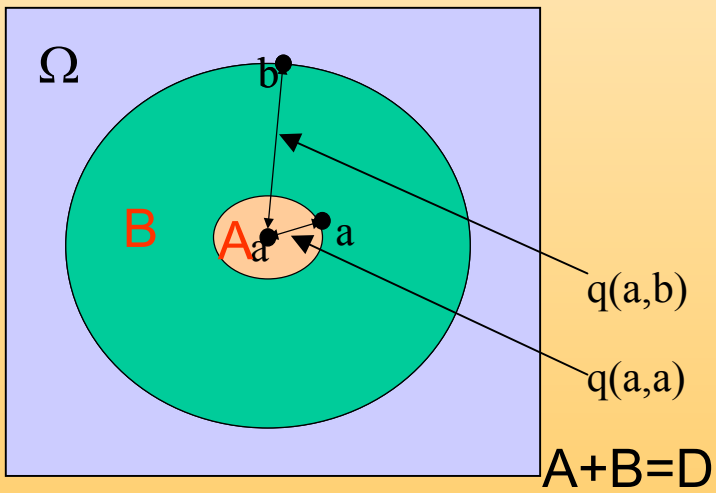


$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)}{P(D)}$$

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

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

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

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

Une nouvelle approche probabiliste (4)

Prise en compte des deux origines:

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

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

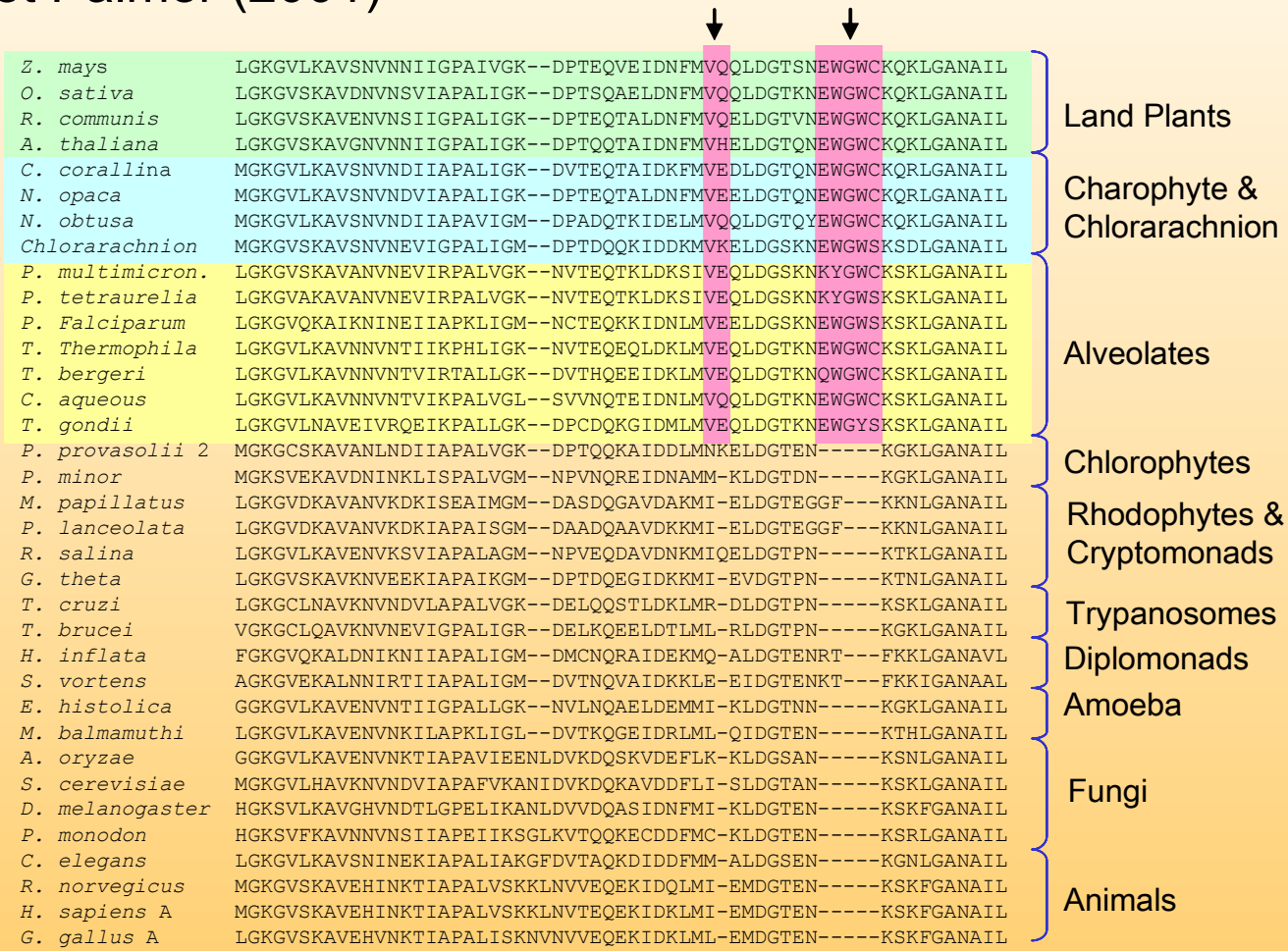
Calcul final de la distance évolutive:

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

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

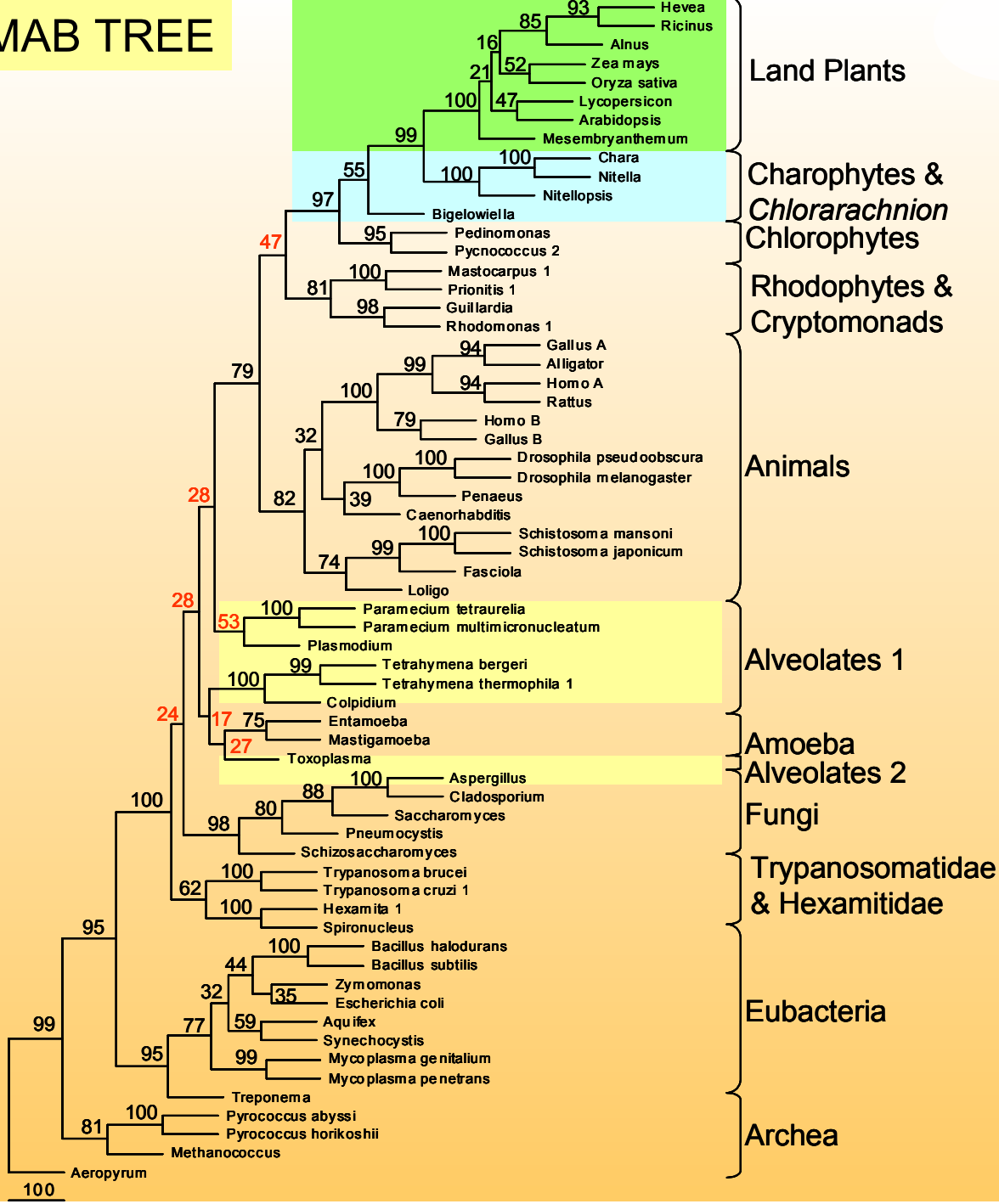
Exemple 2: l'énolase(1)

Keeling et Palmer (2001)



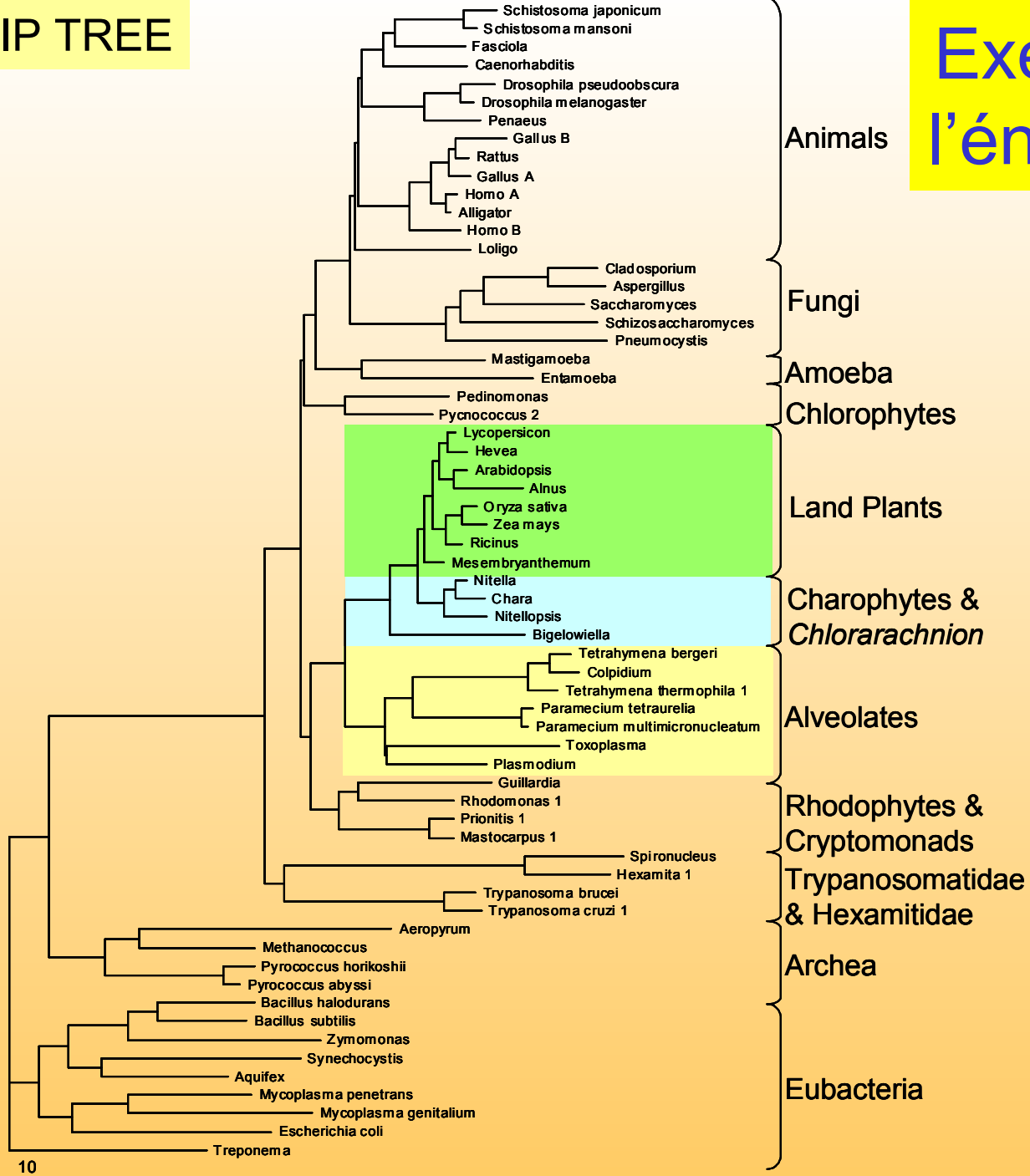
Bastien O., Ortet P., Roy S., and Maréchal E. (2005) BMC Bioinformatics 6:49

Exemple 2: l'énolase(2)



Exemple 2: l'énolase(3)

TULIP TREE



Conclusion générale

- La théorie de l'information fournit un cadre adapté pour reformuler certains principes néo-Darwinien dans des termes mathématiques

Bastien O., Roy S., and Maréchal E. (2005) C.R. Biol. 328:445-453

- Le modèle CSHP permet de construire des arbres phylogénétiques en tenant compte de la totalité de l'information mutuelle du système mais également en tenant compte des taux de mutations par sites

Botte C, Jeanneau C, Snajdrova L, Bastien O, Imberty A, Breton C, Marechal E. (2005) J Biol Chem. 280 (41) 34691-34701

- La majoration du modèle probabiliste par un rapport de Z-Score est très efficace en terme de résultats mais présente 2 inconvénients

1) coûte cher en temps de calcul

2) Reste une majoration. Un modèle plus précis est à l'étude

- Projet d'interface avec P.Ortet (CEA-Cadarache, DEVM)

Perspectives

(i) Amélioration du modèle

- Meilleur approximation du modèle (probabilités d'alignements)

- Prise en compte de la différence de pression évolutive sur les résidus

(ii) Phylogénie à partir de comparaisons pairwise. Application aux transferts horizontaux: Phylogénie moléculaire des protéines de Plasmodium falciparum.

(iii) Exploitation automatique des bases de données d'alignements collectant les z-scores (Teraprot, Decrypthon, CluSTr (EBI)).

Remerciements

Laboratoire de Physiologie Cellulaire Végétale (CEA Grenoble)
Eric Maréchal
Olivier Bastien

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

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

Département d'Écophysiologie Végétale et Microbienne (CEA Cadarache)
Philipe Ortet

Laboratoire Imagerie Médicale quantitative
Sylvain Lespinats
Bernard Fertil

Gene-IT
Jean-Jacques Codani
Karine Métayer