L’origine évolutive de la forme de la distribution des scores d’alignements de séquences biologiques

Olivier Bastien

17 Mai 2011

Plan

- Comparaison de séquences biologiques : les bases

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

- Théorie de la fiabilité et alignement de séquence

- Un nouveau modèle évolutionniste pour la distribution des scores d'alignements

- 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:

$$\begin{cases} q_{ij} \geq p_i\, p_j \Rightarrow S_{ij} \geq 0 \\ q_{ij} \leq p_i\, p_j \Rightarrow S_{ij} \leq 0 \end{cases}$$

Substitution favorable

Substitution défavorable

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 et bien d'autres…)

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

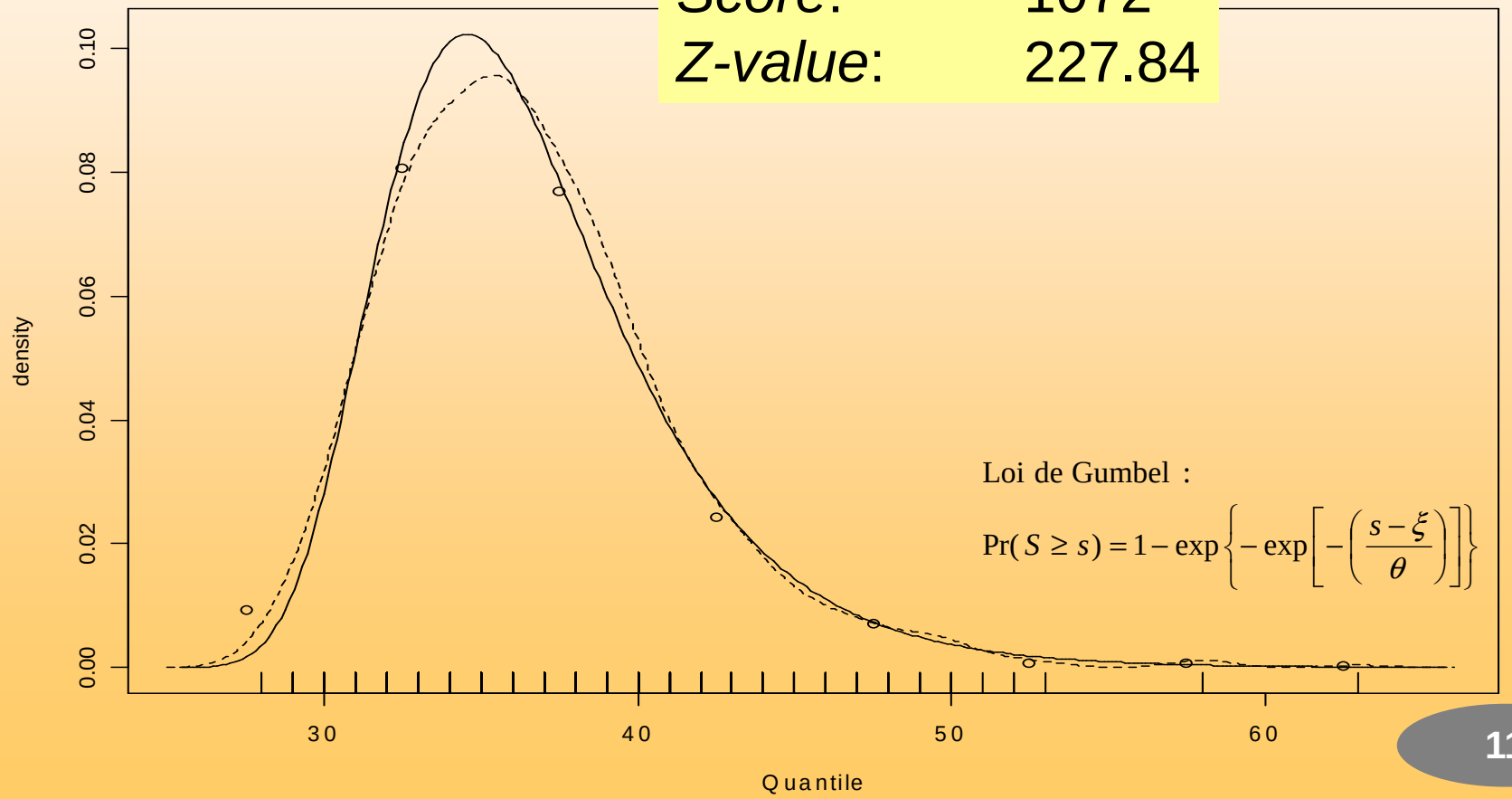
3- On observe la distribution des 1000 $s(a,b^*)$. Où se situe $s(a,b)$ dans cette distribution?

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

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 fournie un majorant pour la probabilité recherchée (la e-value)

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

2) On sait que si S suit une loi de type distribution des valeurs extrêmes

$$P(S(A,B) \leq s) = \exp(-\exp(-\frac{s-\theta}{\beta}))$$

Alors Z suit une loi de type

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

Bastien et al. (2004) Bioinformatics 20 (4): 534-537.

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

1- Quelle est l'origine évolutive de cette loi?

2- Est-elle compatible avec un scénario d'évolution des séquences?

3- Et en particulier: Est-elle compatible avec un mécanisme Darwinien d'évolution-divergence?

Plan

- Comparaison de séquences biologiques : les bases

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

- Le CSHP permet le calcul de distances évolutives

- Théorie de la fiabilité et alignement de séquence

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

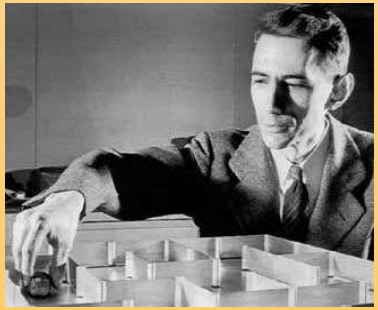
- Conclusion générale

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)

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

$$s(a,b) = I(a;b)$$

Bastien et al. (2005) BMC Bioinformatics 6: 49.
Bastien et al. (2005) C.R. Biol. 328(5): 445-453.

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

Plan

- Comparaison de séquences biologiques : les bases

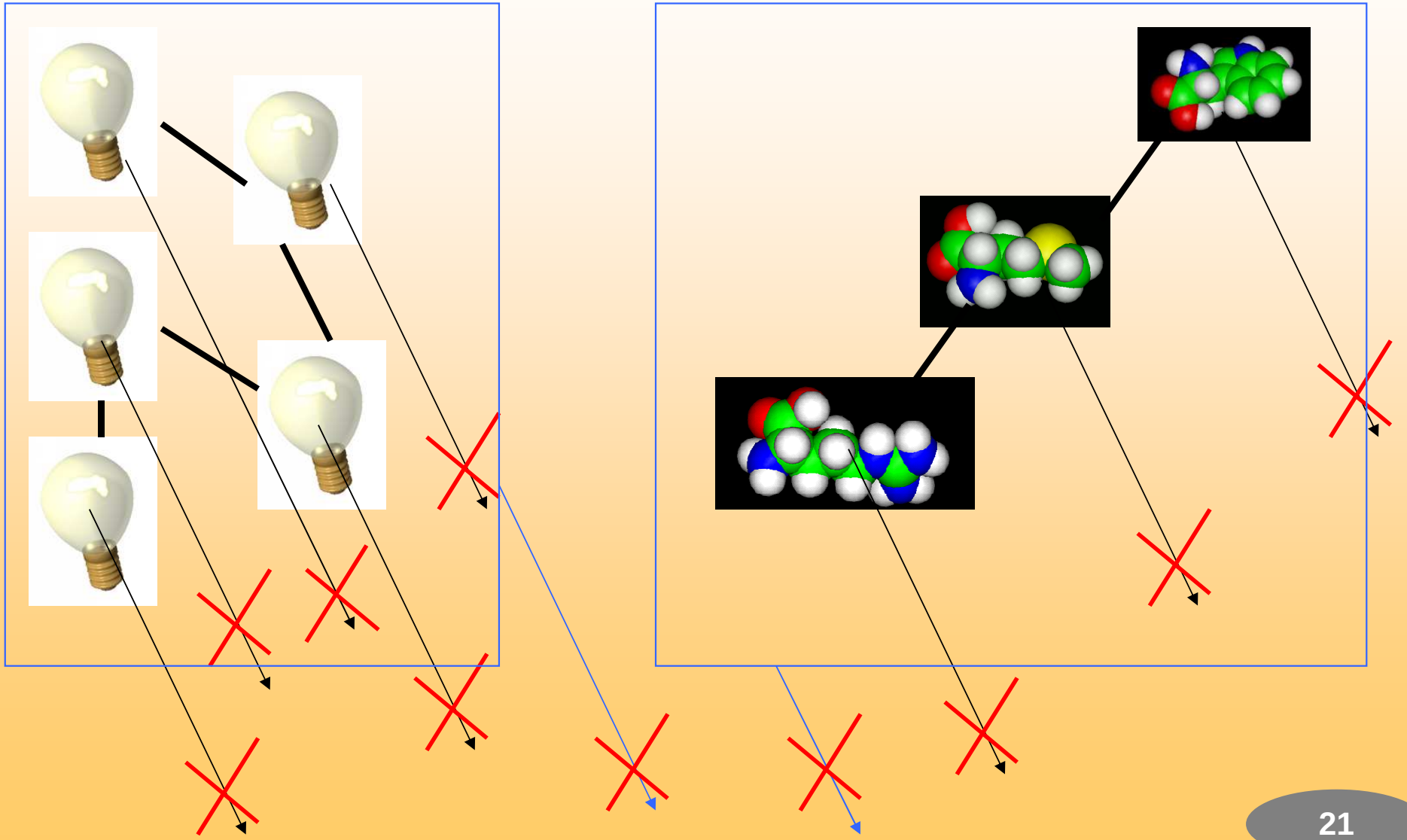
- Comparaison de séquences dans le cadre de la théorie de l'information

- Théorie de la fiabilité et alignement de séquence

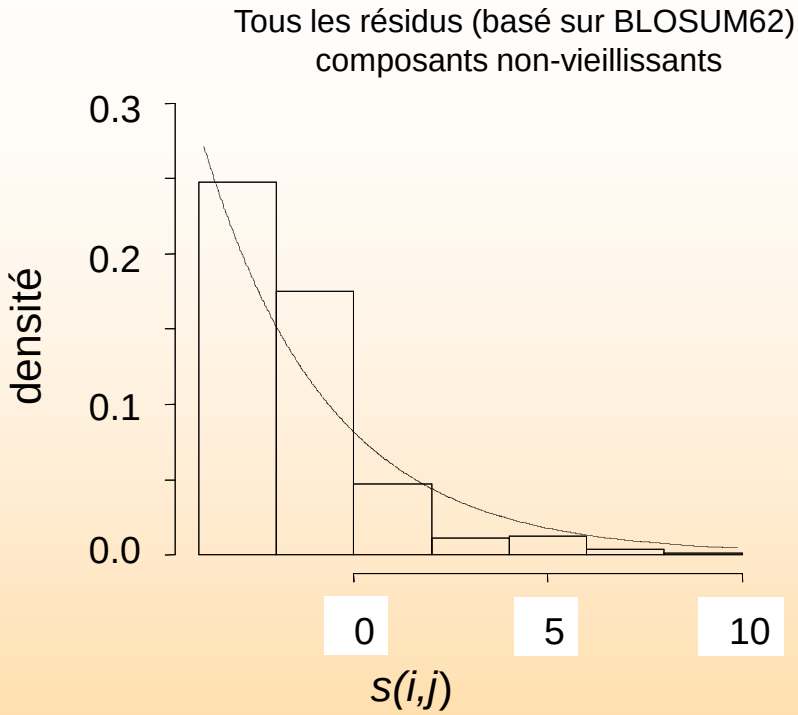
- Un nouveau modèle évolutionniste pour la distribution des scores d'alignements

- Conclusion générale

La théorie de la fiabilité



L’évolution des monomères en fonction du temps



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

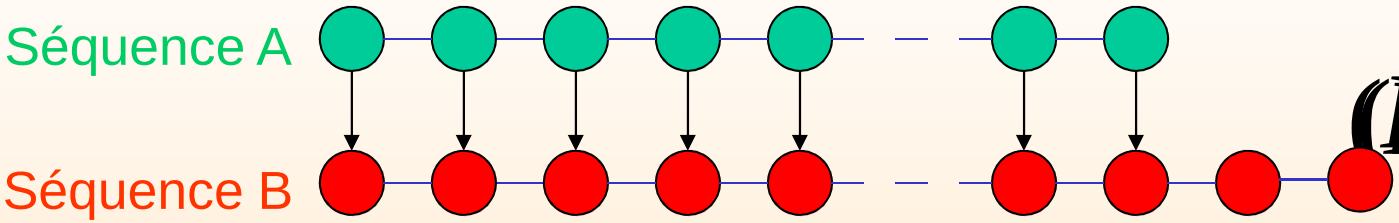
+

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

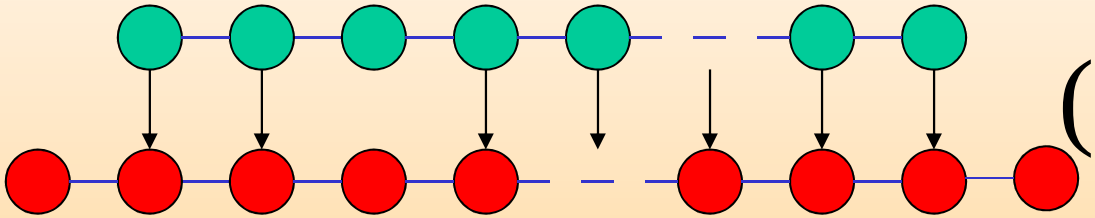
$$\psi(x) = \lim_{dx \to 0} \frac{P(x - dx < X \leq x / X \leq x)}{dx}$$

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

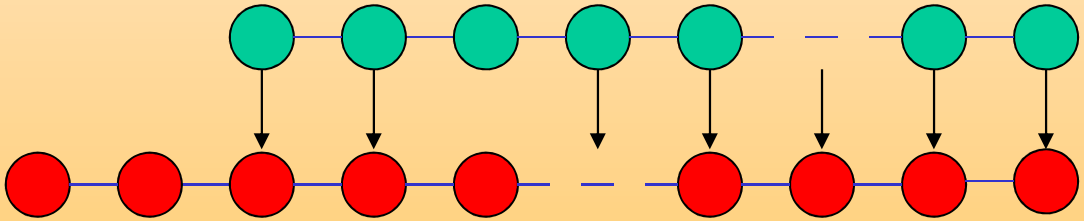
Le calcul de la loi de probabilité (1)



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



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



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

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

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 obtient la distribution de probabilité de la *Z-value*:

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

Bastien O. and Maréchal E.. (2008) BMC Bioinformatics 9: 332.
Bastien, O. (2008) Evolutionnay Bioinformatics 4: 41-45

Conclusion générale en 2006

- La théorie de l'information (couplée à la théorie de la fiabilité qui ajoute l'aspect contrainte sur 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

- On peut dériver la forme de l'équation de Karlin-Altchul à partir d'un processus évolutif primaire.

Mais ce que l'on voudrait: partir d'un processus évolutif d'évolution-spéciation et obtenir un nouveau modèle similaire au modèle de Karlin-Altchul.

Premières observations et motifs d'insatisfactions (1)

- L'approche précédente permet de donner des interprétations théoriques aux paramètres du modèle de Karlin-Altschul

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

Probabilité que les deux séquences conservent la même quantité d'information mutuelle lorsqu'il se produit une mutation dans l'une des deux séquences

Bastien, O. (2008)
Evolutionnay Bioinformatics 4: 41-45

Probabilité que la relation de parenté entre deux séquences cesse d'être détectable par unité de scores

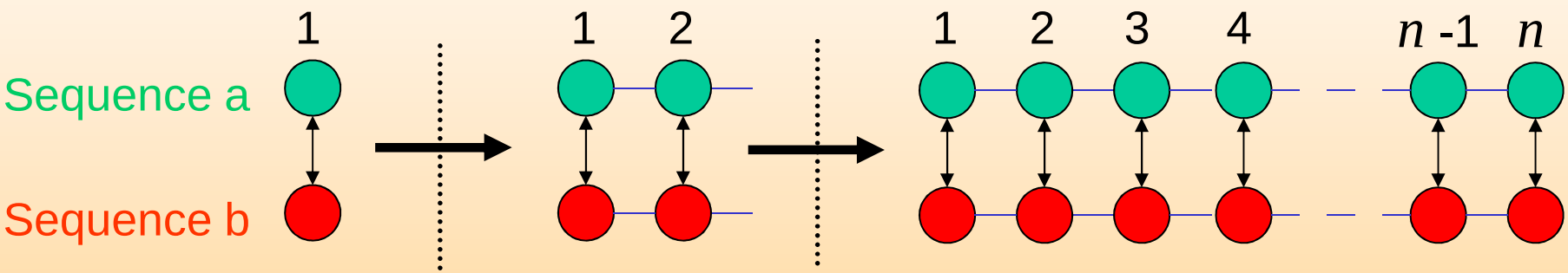
Bastien O. and Maréchal E.. (2008)
BMC Bioinformatics 9: 332.

MAIS…

Premières observations et motifs d'insatisfactions (2)

Tous les procédés précédents de constructions de scores à partir de séquences sont tout sauf évolutionnistes

Impossibilité de partir sur la démarche de K-A



$$S(a,b) = S(a_1,b_1) \qquad S(a,b) = \sum_{i=1}^{2} S(a_i,b_i) \qquad S(a,b) = \sum_{i=1}^{n} S(a_i,b_i)$$

L'évolution de l'information mutuelle entre deux séquences ne se fait certainement pas en alignant les paires de résidues deux-à-deux

Plan

- Comparaison de séquences biologiques : les bases

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

- Théorie de la fiabilité et alignement de séquence

- Un nouveau modèle évolutionniste pour la distribution des scores d'alignements

- Conclusion générale

Les idées de base (1)

1- construire un modèle d'évolution des séquences comprenant:

- Le phénomène de duplication

  les gènes se dupliquent dans les mêmes espèces

  la spéciation implique la création de nouveaux gènes semblables au début aux précédents

- Le phénomène de divergence

  elle est due en grande partie au fait que les macromolécules biologiques sont des entités de grandes dimensions

2- prendre en compte le fait que la forme qualitative de la loi de distribution des scores est la même quelle que soit la séquence requête, la séquence sujette (base de données?) et quelle que soit l'époque où on effectue la comparaison.

Les idées de base (2)

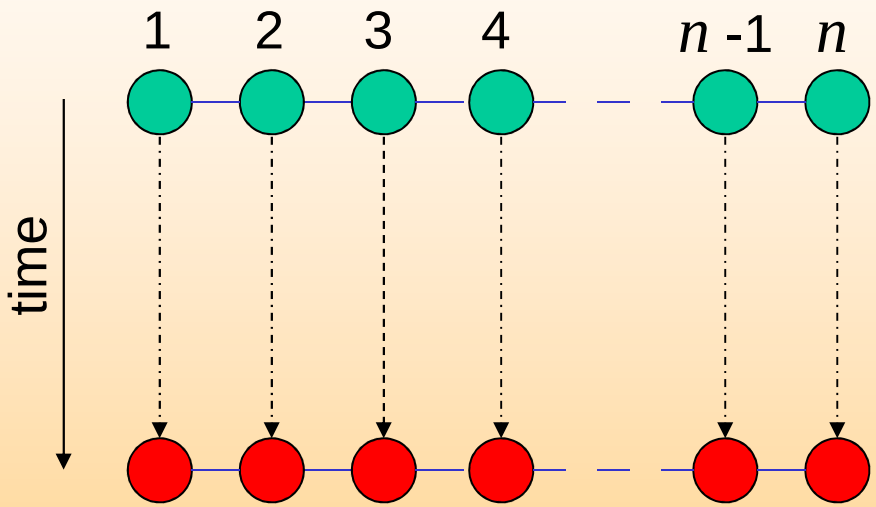
3- on va partir d'une seule séquence au temps t=0.

4- par unité de temps, toutes les séquences présentes dans le processus vont diverger de la séquence ancestrale

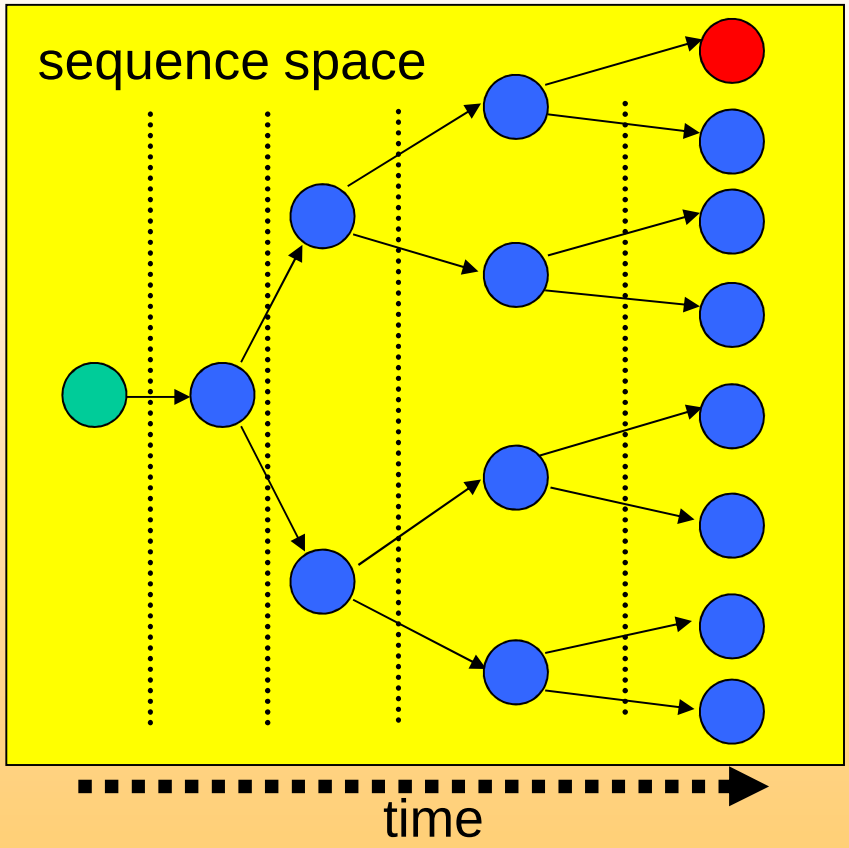
5- par unité de temps, toutes les séquences présentes dans le processus vont se dupliquer

Pour résumer…

Les idées de base (3)



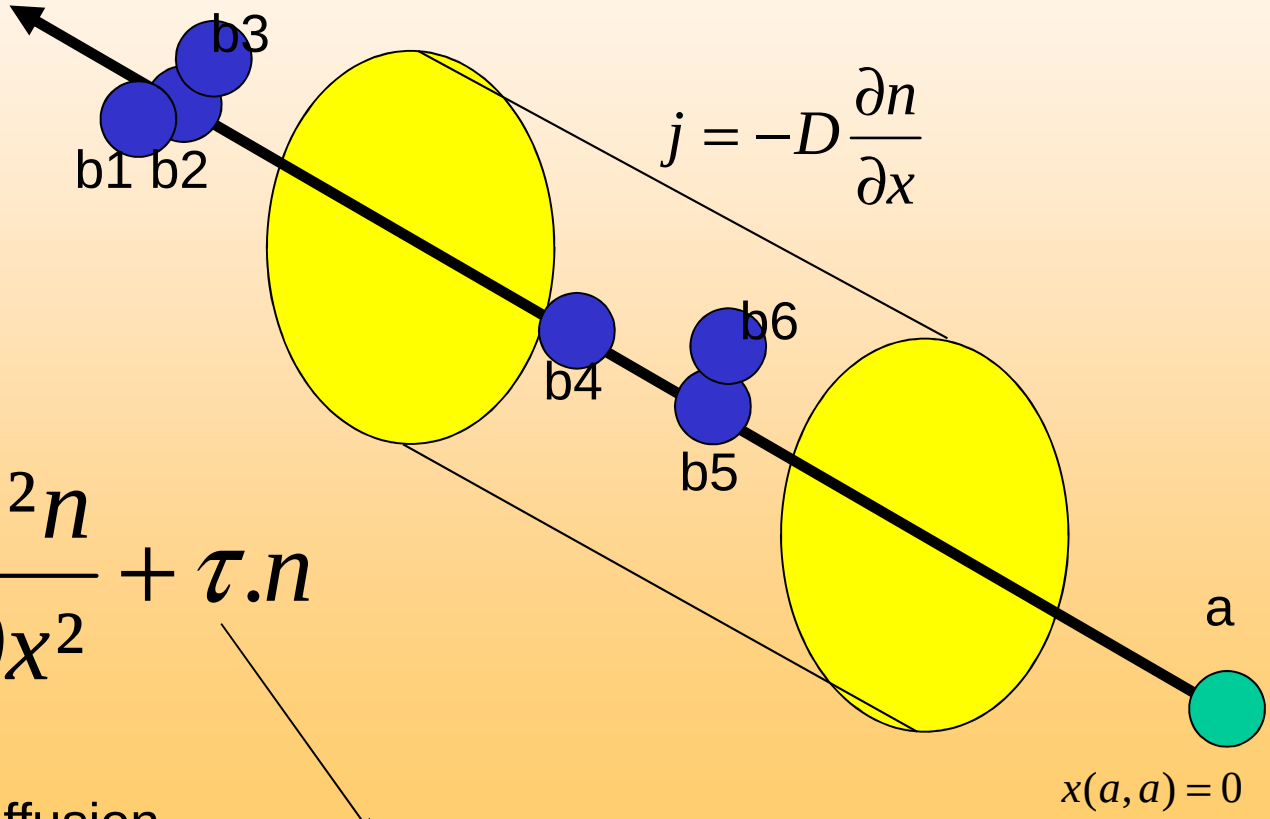
$S(a,b) = f(time) = I(a;b)$, i.e. global measure



$S(a,b) = f(time) = I(a;b)$, i.e. global measure

Première étape: Duplication-divergence des protéines dans l'espace des séquences

- Dans un premier temps, on va étudier la répartition des séquences bi en fonction de leurs distances génétiques x(a,bi) à la séquence ancestrale a



$$\frac{\partial n}{\partial t} = D\frac{\partial^2 n}{\partial x^2} + \tau.n$$

Constante de diffusion
Ici: la divergence

Taux de duplication

Deuxième étape: recherche de solutions indépendantes du temps (1)

- La distance génétique est une distance qui va de 0 à 1, par exemple le pourcentage de résidus différents entre deux séquences.

$$D\frac{\partial^2 n}{\partial x^2}+\tau=0 \qquad \begin{cases} n(0)=0 \\ n(M)=0 \end{cases}$$

Il s'agit d'un problème régulier de Sturm-Liouville dont les solutions sont:

$$n(x)=A\cos(\sqrt{\frac{\tau}{D}}x)+B\sin(\sqrt{\frac{\tau}{D}}x)$$

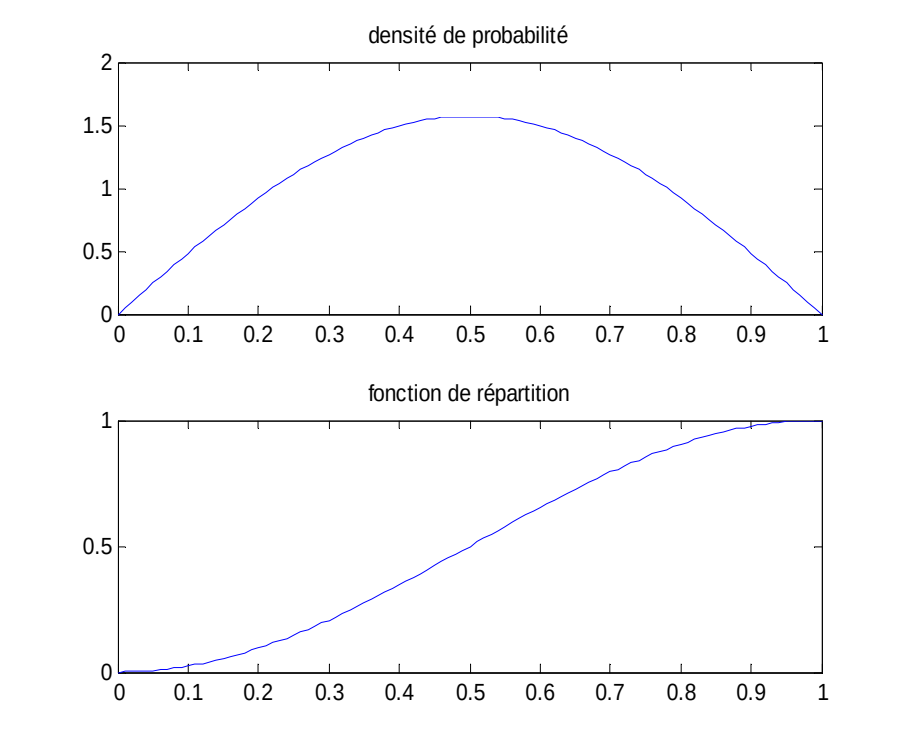
Et donc, en tenant compte des conditions aux frontières:

$$n(x)=B\sin(\frac{l\pi}{M}x)$$

Deuxième étape: recherche de solutions indépendantes du temps (2)

- n(x) est le nombre de séquences présentes entre x et x+dx. En intégrant, on a le nombre de séquences présentes dans tout l'espace.

- Si on normalise à 1, on a la distribution de probabilité des séquences en fonction de la distance génétique



$$\rho(x) = \frac{\pi}{2}\sin(\pi x)$$

$$P(X \leq x) = \frac{1}{2}[1 - \cos(\pi x)]$$

MAIS…

Troisième étape: De la distance génétique à l'information mutuelle entre séquences (1)

- On a la distribution en fonction de la distance génétique x, mais tout ce que l'on sait mesurer, c'est l'information mutuelle I(a,b) entre les séquences

- Il faut donc trouver une relation plausible entre x et I, c'est à dire $\frac{dx}{dI} = f(x, I)$

- Première étape: on effectue un développement de Taylor d'ordre 2 au voisinage de (1,0) (distance génétique maximum, Information mutuelle minimum):

$$f(1-x,0+I) \approx f(1,0) - \frac{\partial f}{\partial x}(1,0)x + \frac{\partial f}{\partial I}(1,0)I - \frac{\partial^2 f}{\partial x \partial I}(1,0)xI + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(1,0)x^2 + \frac{1}{2}\frac{\partial^2 f}{\partial I^2}(1,0)I^2 + \ldots$$

Troisième étape: De la distance génétique à l'information mutuelle entre séquences (2)

- On fait l'hypothèse que $\frac{\partial f}{\partial x}(1,0), \frac{\partial f}{\partial I}(1,0)$, $\frac{\partial^2 f}{\partial x^2}(1,0)$ et $\frac{\partial^2 f}{\partial I^2}(1,0)$ sont nulles au voisinage de (1,0).

- On obtient alors une relation entre x et I sous la forme

$$dx = -\alpha x I dI \quad , \quad \alpha = -\frac{\partial^2 f}{\partial x \partial I}(1,0)$$

- Ce qui conduit à

$$x(I) = \exp(-\alpha(I^2 - \xi^2)$$

Quatrième étape: une nouvelle distribution de probabilité

- On remplace $x(I) = \exp(-\alpha(I^2 - \xi^2)$ dans $\rho(x) = \frac{\pi}{2}\sin(\pi x)$

- Ce qui conduit à une nouvelle distribution de probabilité

Densité de probabilité

$$\rho(s) = \frac{\alpha\pi}{2} s \exp(-\alpha(s^2 - \zeta^2))\sin(\pi \exp(-\alpha(s^2 - \zeta^2))) \quad , \quad s \in [\zeta, +\infty[$$

Fonction de répartition

$$P(S \le s) = \frac{1}{2}[1 + \cos(\pi \exp(-\alpha(s^2 - \zeta^2)))] \quad , \quad s \in [\zeta, +\infty[$$

Cette nouvelle loi de probabilité est une loi à deux paramètres.

une version générale de la nouvelle distribution de probabilité

Densité de probabilité

$$\rho(s) = \frac{\alpha\eta\pi}{2} s^{\eta-1} \exp(-\alpha(s^{\eta} - \zeta^{\eta})) \sin(\pi \exp(-\alpha(s^{\eta} - \zeta^{\eta}))), \quad s \in [\zeta, +\infty[$$

Fonction de répartition

$$P(S \leq s) = \frac{1}{2}[1 + \cos(\pi \exp(-\alpha(s^{\eta} - \zeta^{\eta})))], \qquad s \in [\zeta, +\infty[$$

Cette nouvelle loi de probabilité est une loi à trois paramètres.

Ortet, P. and Bastien, O. (2010) Evolutionary Bioinformatics 6: 159.

Comparaison du nouveau modele avec deux modèles concurrents

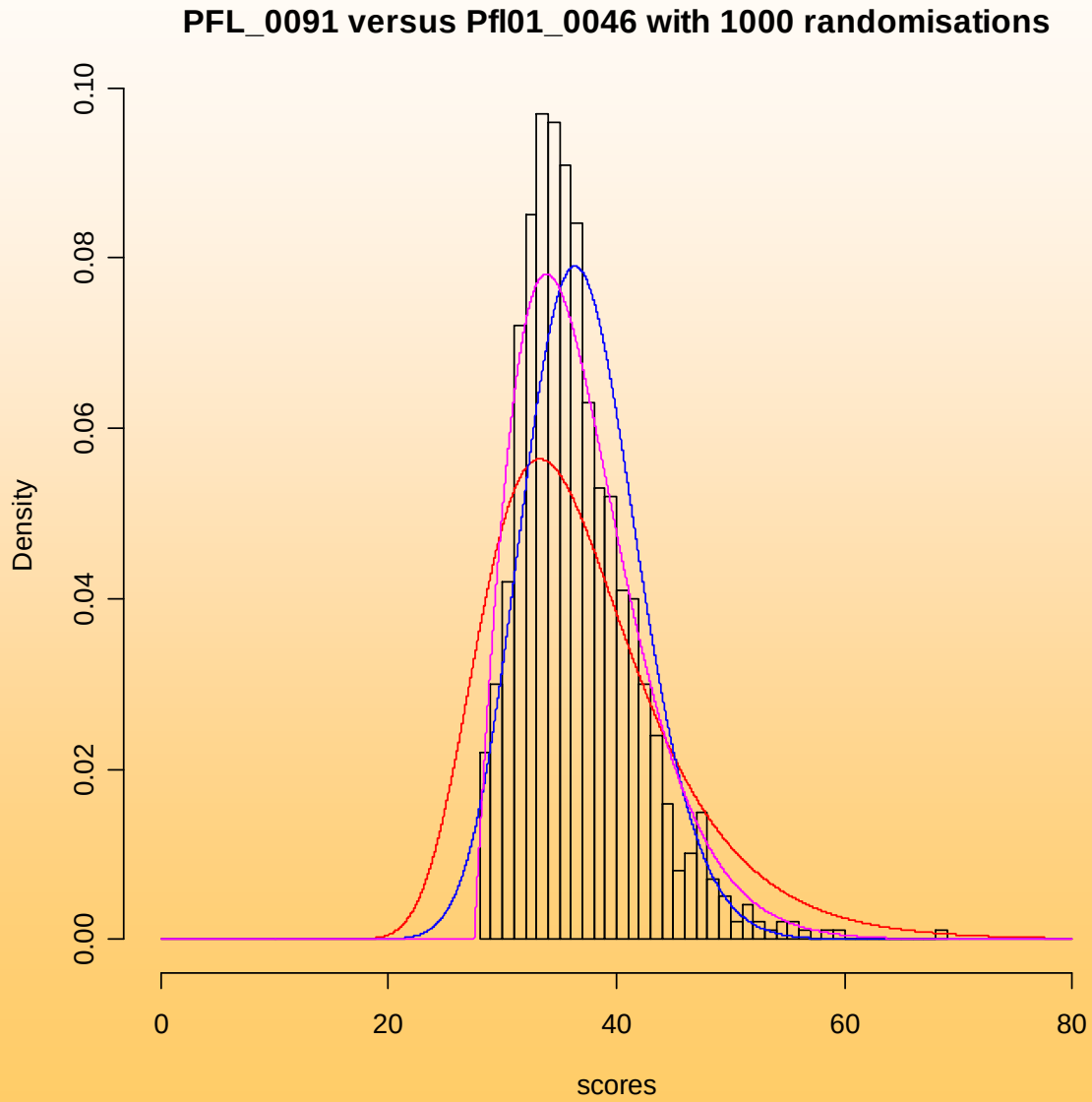
- Le distribution de Gumbel est une distribution à 2 paramètres:

$$p(s) = \frac{1}{\beta}\exp\left(-\frac{(s-\theta)}{\beta}\right).\exp(-\exp\left(-\frac{(s-\theta)}{\beta}\right))$$

- La distribution gamma est une distribution à 3 paramètres:

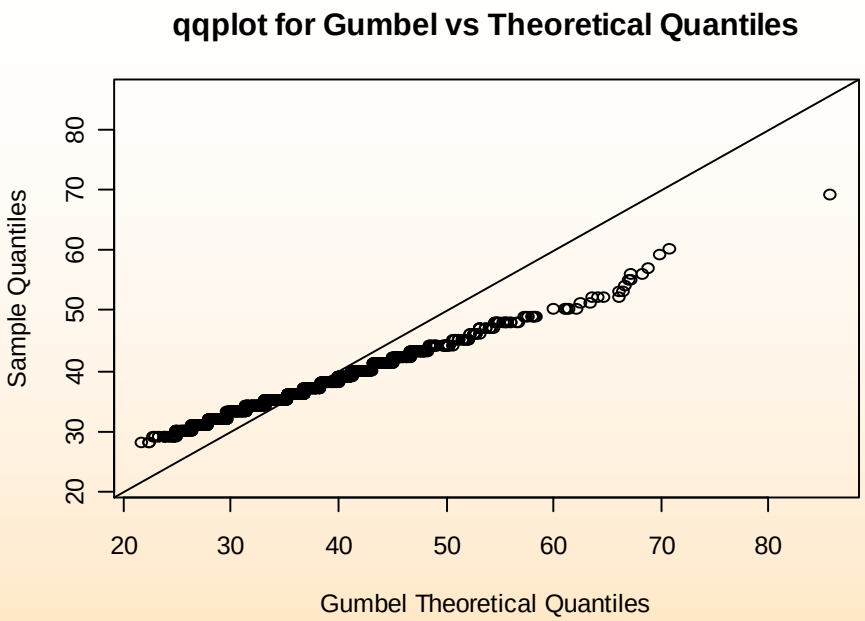
$$p(s) = s^{\delta-1}.\exp(-\frac{s}{\omega}).\left(\Gamma(\delta).\omega^{\delta}\right)^{-1}$$

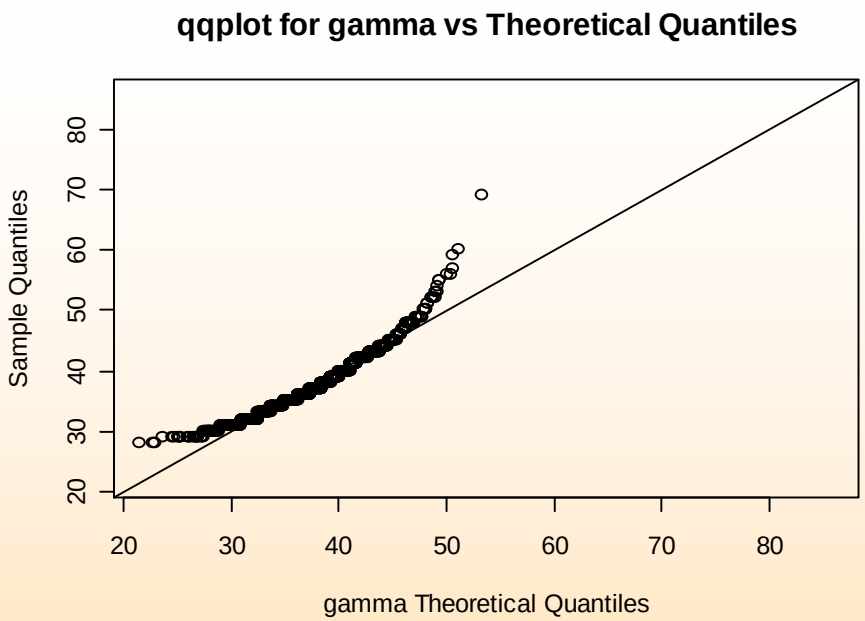
Response Regulator NtrC family proteins in Pseudomonas fluorescens

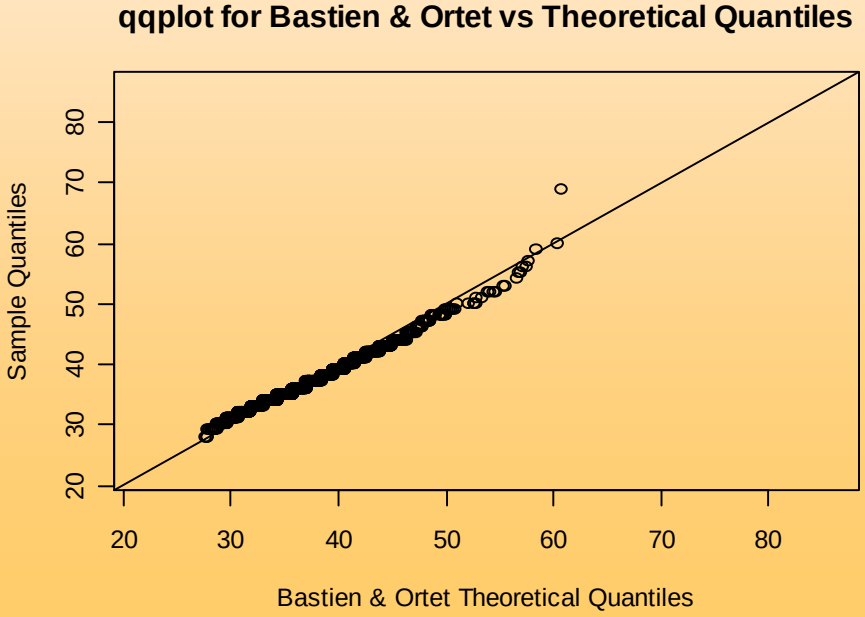


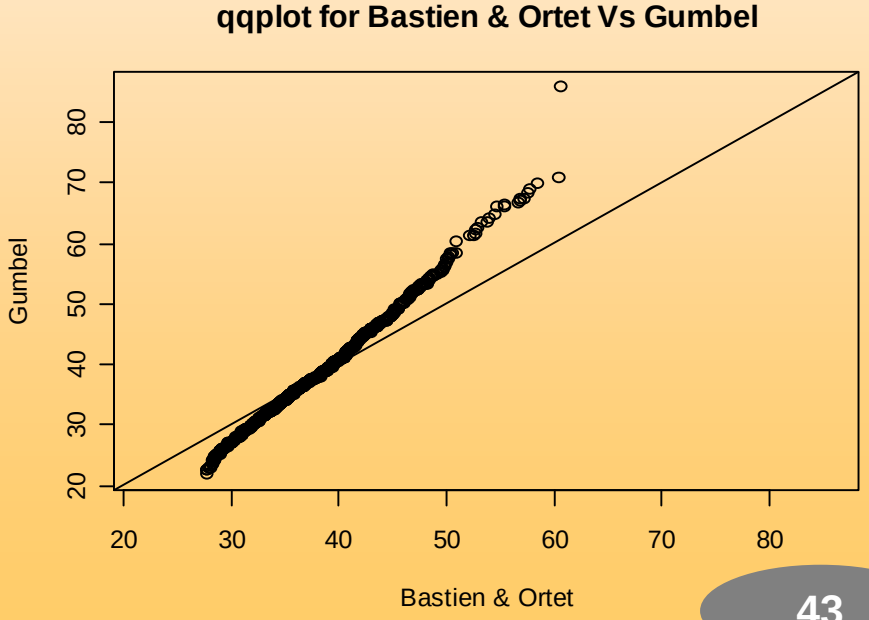
Response Regulator NtrC family proteins in Pseudomonas fluorescens Pf-5 and his homologous proteins in Pseudomonas fluorescens Pf0-1 (Accession numbers PFL_0091 and Pfl01_0046). Only the second sequence was shuffled 1000 times.

Red curve: Gumbel distribution
θ = 33.27876
β = 6.523116.
Blue curve: gamma distribution
δ = 53.04861
ω = 0.6983029.
Purple curve: our model
α=0.001281424
ζ=27.500000245
η=1.999992822.


qqplot for Gumbel vs Theoretical Quantiles
Sample Quantiles
Gumbel Theoretical Quantiles
20
30
40
50
60
70
80

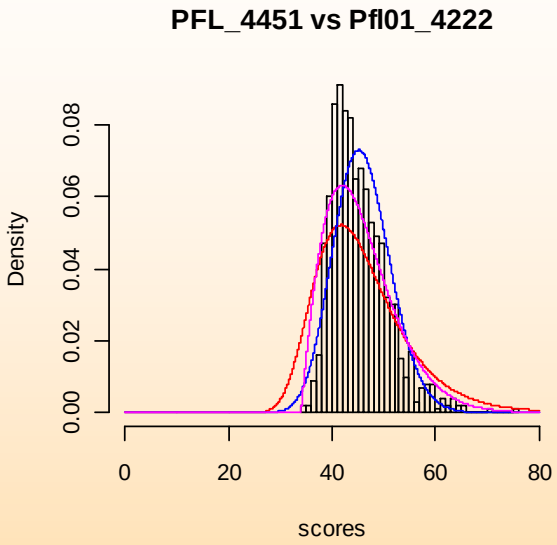

qqplot for gamma vs Theoretical Quantiles
Sample Quantiles
gamma Theoretical Quantiles
20
30
40
50
60
70
80

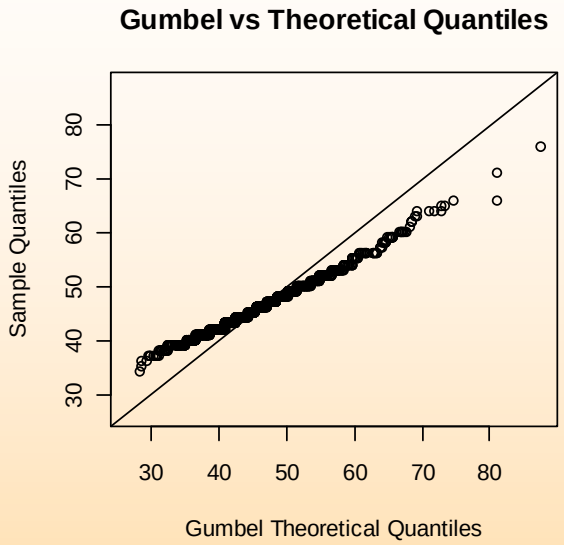

qqplot for Bastien & Ortet vs Theoretical Quantiles
Sample Quantiles
Bastien & Ortet Theoretical Quantiles
20
30
40
50
60
70
80


qqplot for Bastien & Ortet Vs Gumbel
Gumbel
Bastien & Ortet
20
30
40
50
60
70
80

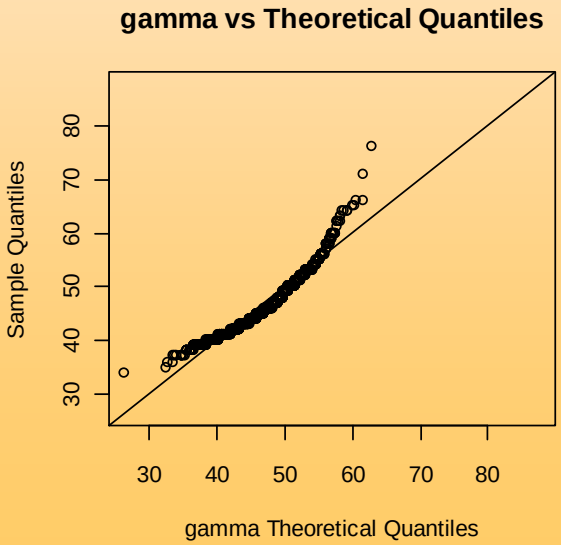
Two-component system, NarL family, sensor histidine kinase Regulator NtrC family proteins

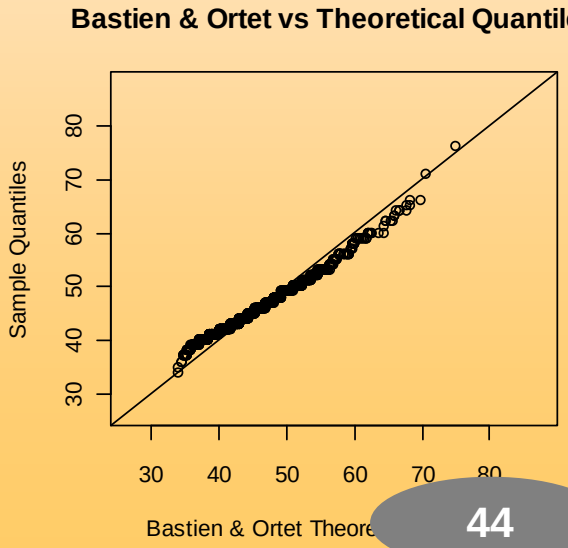
two-component system, NarL family, sensor histidine kinase in Pseudomonas fluorescens Pf-5 and his homologous proteins in Pseudomonas fluorescens Pf0-1 (Accession numberAs PFL_4451 and Pfl01_4222).





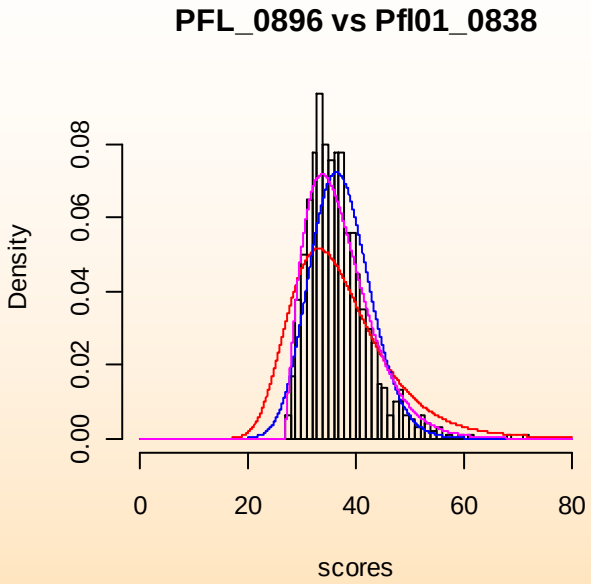
Red curve: Gumbel distribution
θ = 41.79989
β = 7.047815.
Blue curve: gamma distribution
δ = 69.67213
ω = 0.6583407.
Purple curve: our model
α=0.0006687308
ζ=33.99296
η=2.053335.

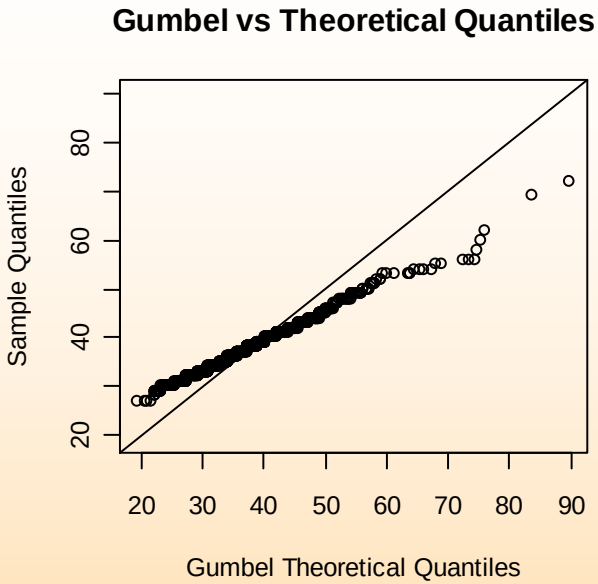




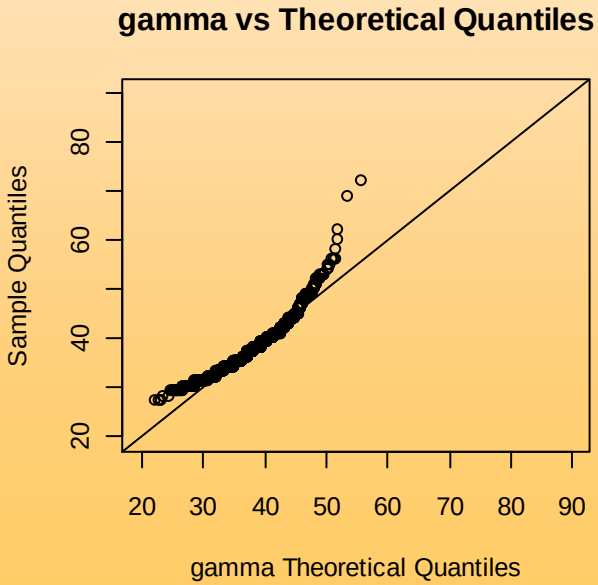
rod shape-determining protein MreB

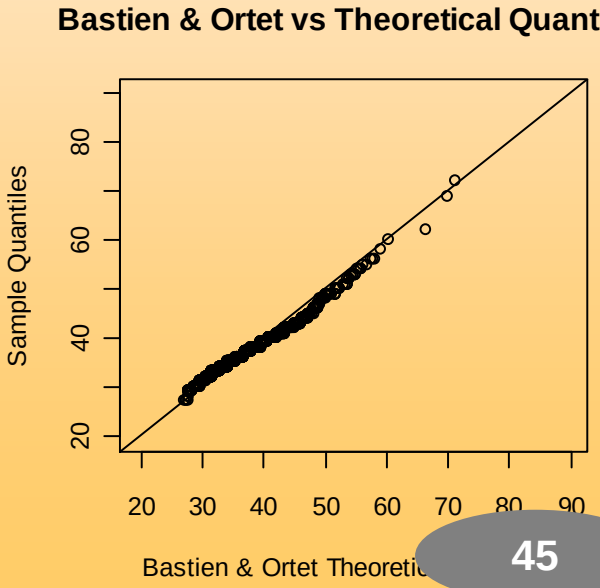
rod shape-determining protein MreB in Pseudomonas fluorescens Pf-5 and his homologous proteins in Pseudomonas fluorescens Pf0-1 (Accession numbers PFL_0896 and Pfl01_0838).





Red curve: Gumbel distribution
θ = 33.20788
β = 7.1206
Blue curve: gamma distribution
δ = 45.1806
ω = 0.825974.
Purple curve: our model
α=0.001707993
ζ=26.95872
η=1.908571.





Remerciements

Laboratoire de Physiologie Cellulaire Végétale (CEA Grenoble)
Eric Maréchal
Sylvaine Roy

CEA Saclay - LIST Laboratoire Intelligence Multi-capteurs et Apprentissage
Sylvain Lespinats

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

TIMC-TIMB (Université Joseph Fourier)
Nicolas Glade

Adaptation and Pathogenesis of Microorganisms (LAPM)
Mohamed Ali Hakimi