

OPEN ACCESS
Full open access to this and thousands of other papers at <http://www.la-press.com>.

Where Does the Alignment Score Distribution Shape Come from?

Philippe Ortet¹ and Olivier Bastien²

¹CNRS (UMR 6191)-CEA Cadarache-Aix-Marseille Université, Laboratoire d'Ecologie Microbienne de la Rhizosphere, Institut de Biologie Environnementale et Biotechnologie, CEA Cadarache, F-13108 Saint Paul-lez-Durance, France.

²UMR 5168 CNRS-CEA-INRA Université J. Fourier, Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France.

Corresponding author email: olivier.bastien@cea.fr

Abstract: Alignment algorithms are powerful tools for searching for homologous proteins in databases, providing a score for each sequence present in the database. It has been well known for 20 years that the shape of the score distribution looks like an extreme value distribution. The extremely large number of times biologists face this class of distributions raises the question of the evolutionary origin of this probability law.

We investigated the possibility of deriving the main properties of sequence alignment score distributions from a basic evolutionary process: a duplication-divergence protein evolution process in a sequence space. Firstly, the distribution of sequences in this space was defined with respect to the genetic distance between sequences. Secondly, we derived a basic relation between the genetic distance and the alignment score. We obtained a novel score probability distribution which is qualitatively very similar to that of Karlin-Altschul but performing better than all other previous model.

Keywords: sequence alignment scores, Karlin-Altschul theorem, mutual information, sequence space

Evolutionary Bioinformatics 2010:6 159–187

doi: [10.4137/EBO.S5875](https://doi.org/10.4137/EBO.S5875)

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.



Introduction

Comparison of biological macromolecules by the means of sequence alignments has become an everyday task for biologists, for extremely diverse purposes such as genomic sequencing, structural modelling, functional inference, phylogenetic reconstruction, etc. Comparison methods rely on a fundamental postulate that one can simply state as: “the closer in evolution, the more alike and conversely, the more alike, probably the closer in evolution”. Alignment of two sequences is typically done by maximizing a given quantity, called the score, which reflects the shared features of the two biological entities.^{1–3} Scoring matrices are used to maximize the summed scores of compared residues and find optimal local alignments, computed with a dynamic programming procedure.^{1–3} Scoring matrices have been found^{4–6} to be similarity matrices which can be derived from previously known alignments, and to be of the form $s(i, j) \propto \log_2(q_{ij}/p_i p_j)$ where i and j are aligned amino acids, q_{ij} the frequency of the observation: “ i is aligned with j ”, (the target frequency), p_i and p_j are respectively the frequency of i and j , ie, the background frequency. From an information theory point of view, scores between residues can be considered as estimations of the mutual information between the elementary events called “amino acids”.⁷ The average value H of the substitution matrix, $H = \sum_{i,j} q_{ij} \log_2(q_{ij}/p_i p_j)$, is the Kullback-Leibler distance between the observed joint distribution of the amino acids i in a sequence a and the amino acids j in a homologous sequence b . H is also the mutual information between the random variable A: “drawing an amino acids i from the sequence a at a given position x ” and the random variable B: “drawing an amino acids j from the homologous sequence b at the homologous position to x ”. H is called the relative entropy of the substitution matrix.⁸ It is a measure of the dependence between the two random variables A and B in the homologous sequences which have served to construct the substitution matrix.

Confidence in pairwise alignments of biological sequences, obtained by various methods such as Blast^{1,2} or Smith-Waterman,³ is critical for automatic analyses of genomic data. Since the degree of similarity is usually assessed by the sequence alignment score, it is necessary to know if a score is high enough to indicate a biologically interesting alignment.⁹ In the asymptotic limit of long sequences,

the Karlin-Altschul model¹⁰ computes a P -value assuming that the number of high scoring matching regions above a threshold is Poisson distributed. This model expresses the probability of having a random sequence with a score (called the random score) S less or equal to the observed score as:

$$P(S \leq s) = \exp(-kmn \exp(-\lambda s)) \quad (1)$$

where m and n are the length of the two compared sequences and k and λ are two parameters depending of the sequence compositions which must be adjusted to the data.^{11,12} The P -value, which is a measure of the statistical significance of the score, is obtained from this equation by considering that $P\text{-value} = 1 - P(S \leq s) = P(S > s)$. It is the probability of having a random sequence with a score S larger than the observed score s . This particular distribution (1) is known as the extreme values distribution of type I, or Gumbel distribution, and characterises the distribution of extreme events like rogue waves colliding on a sea wall, large wildfires and, among others, maximal sequence scores. Karlin and Altschul derived this distribution by considering a random sequence alignment score as an addition of random residue pairs’ scores taken from a substitution matrix^{4–7,10} (see Fig. 1A) and studied the behaviour of such scores when using an alignment algorithm (which often belongs to the dynamic programming algorithm class^{1–3}). This extremely important result is well established for ungapped pairwise sequence alignments and is strongly suspected to be applicable not only for a large class of gapped pairwise alignments,¹² (depending on the penalty for gap opening and extension)⁹ but also for a large class of alignments such as hmm profiles.^{13,14} This last point suggests that this distribution (or qualitatively similar distributions) seems to play a key role in computational biology.^{12–16} Comet et al¹⁷ showed that the distribution of scores coming from the comparison of one sequence versus a database display the same qualitative distribution (a Gumbel-like distribution) as the comparison of random sequences with a biological sequence. From a theoretical evolutionary biology point of view, the link between the distribution of scores between homologous proteins and the processes which have led to these proteins is not clear. Indeed, as mentioned in the following paragraph,

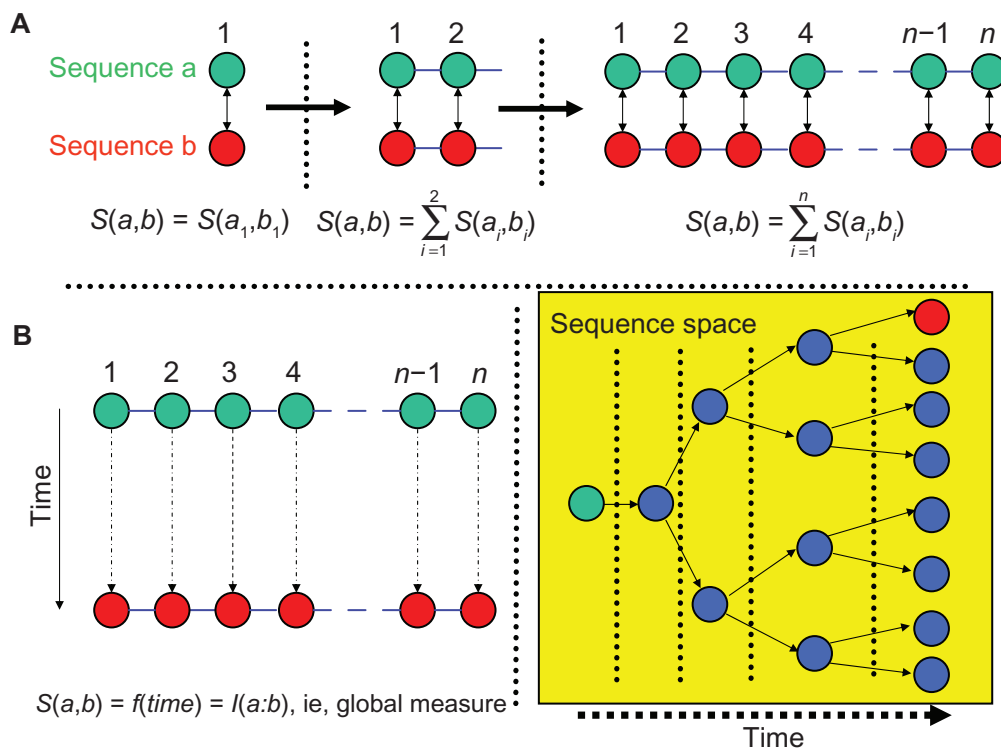


Figure 1. A) In classical alignment score statistical inference, random scores are computed by adding random pairs of aligned residues and then, summing all along the random alignment. **B)** In the new model, all sequences present in the sequence space diverge from one another. The score is a global measure of the shared amount of information between pairs of sequences which decreases with time.

scores between proteins are a measure of the information that each of them possess regarding the others. As a consequence, it should be possible to construct a model reflecting the fact that the observed distribution of scores is a consequence of the process of protein evolution. Unfortunately, the Karlin-Altshul model is not an evolutionary one, as it can be seen in Figure 1A. This is due to the fact that homologous proteins did not evolve by adding residues pairs to aligned sequences.

Recently, it has been demonstrated that the Karlin-Altshul model can be derived with no reference to the extreme events theory, using a simple approach combined with recent results in reliability theory.^{18,19} Sequences were considered as systems in which components are amino acids. As a consequence, these systems have a high redundancy of information reflected by their alignment scores. Evolution of the information shared between aligned components determines the Shared Amount of Information (SAI) between sequences, ie, the score. Then, the Gumbel distribution parameters λ and k of aligned sequence scores finds a theoretical rationale. The first, λ , is the Hazard Rate of the distribution of scores between residues¹⁹

and the second, k , is the probability that two aligned residues do not lose bits of information (ie, conserve an initial pairing score) when a mutation occurs.¹⁸ This result also suggests that alignment score distributions could result from a purely evolutionary process. The question remains why this particular class of distribution seems to be widely spread in nature, especially in biological sequence comparisons.

Here, we are making the hypothesis that the over-representation of this class of probability distributions when comparing biological sequences is a consequence of both the nature of the process by which all these sequences arose, the speciation-divergence process, and of the nature of the measure used to compare these sequences, ie, the SAI, an information theory mathematical object. In order to test this hypothesis, we will consider the SAI as a global measure between sequences lying in a sequence space (see Fig. 1B). Starting from a unique sequence, we will consider the distribution of all sequences present in this space after having applied a basic duplication-divergence evolutionary process and after a sufficiently long time.

If the SAI is global measure between sequences during the evolution process, it is clear that the



following model should apply to score providing by local alignment. Indeed, local alignments focus on segment pairs that maximize alignment score.¹⁻³ Part of the homologous sequences which are aligned in the local alignment are often viewed as homologous segments, whereas the rest of the sequences can have been obtained by adding different biological units called domains. For example, local alignments are used to derive amino acids substitution matrices.^{5,6}

Results

Duplication-divergence protein in sequence space

If one considers a sequence space, like the CSHP⁷ (the Configuration Space of Homologous Proteins is a sequence space where all proteins are identified relative to a particular one) with a sequence a_{ref} as reference, all other sequences b can be placed in this space and their evolution can be studied by means of their genetic distance x (evolutionary time) from the sequence a_{ref} or by the mean of their mutual information (ie, the score $s(a_{ref}, b)$) with a_{ref} . Let us begin by considering evolution with respect to genetic distance. This distance lies between 0 and 1 (for example, x can be the percentage of different residues between two aligned sequences).²⁰

Now, consider the simple evolutionary process where there is just one sequence a_{ref} at the beginning of the process. Per unit time, all existing sequences can produce other sequences at a constant rate τ (this construction is close to the so-called molecular clock hypothesis). We can consider that τ takes into account the rate of loss of sequences if this is also a constant. In the model, τ plays the role of speciation, or duplication, rate. In addition to the duplication rate, all sequences produced in the CSHP diverge from others as time goes on. In particular, they diverge from a_{ref} . This fact is largely a consequence that in a high dimensional space (a protein of length n can be considered as an n -dimensional object in which each vector components can take 20 values), the probability to return to the initial state tends to 0 as n tends to infinity. This observation may be related to the study of random walks on Z^d (the Euclidian product of the set of relative integers d times) and on the seminal work of Polya on the relation between the probability to return in an initial state and the dimension of the space.^{21,22} From the point of view of the genetic

distance x , this process can be viewed as a diffusion process along the interval $[0,1]$, and we can make as a first approximation an analogy with the first Fick law of diffusion.²³ We can write the diffusion flux (which is a flux of molecules from the point $x = 0$ to the point $x = 1$) to be proportional to the sequence concentration gradient, that is to say $j = -D \partial n / \partial x$, where n is the concentration of sequences, x is the genetic distance and D the diffusion coefficient in the sequence space (dimension $x^2 t^{-1}$). Applying the law of conservation of the number of sequences in addition to the production of them, we arrive to the final classical equation

$$\frac{\partial n}{\partial t} = D \frac{\partial^2 n}{\partial x^2} + \tau n \quad (2)$$

where τ has a dimension t^{-1} and where n has to be determined on the $[0,1]$ interval of the x axis. Indeed, solutions outside this interval will have no biological meaning even if these solutions could be determined under some conditions.

Stationary solutions

Considering that there seems to exist a general probability distribution class for sequence comparisons scores⁹⁻¹⁹ and that it seems (not the distribution parameters but only the qualitative shape, ie, the class of the distribution) to be independent of the sequences, and hence from the time since their divergence, it is natural to search a time independent solution to the equation (2) with the natural boundary conditions $\{n(0) = 0, n(M) = 0$ where M is the maximum of the genetic distance (here $M = 1$), and to verify with data if this solution can be applied to our problem. Remembering that solutions outside this interval will have no biological meaning, there is no need to impose $n(x) = 0$ for $x \notin [0,1]$. As a consequence, there is no need to impose the conditions $\{\partial n / \partial x(0) = 0, \partial n / \partial x(M) = 0$. These remarks and the equation (2) lead to the time independent equation to solve

$$D \frac{\partial^2 n}{\partial x^2} + \tau n = 0, n(0) = 0, n(M) = 0 \quad (3)$$

This classical problem in partial differential equation theory is known as the Regular Sturm-Liouville problem^{24,25} and the general solution of the stationary equation is known to be



$n(x) = A \cos(\sqrt{\tau/D}x) + B \sin(\sqrt{\tau/D}x)$ which becomes, after considering boundary conditions

$$n(x) = B \sin\left(\frac{l\pi}{M}x\right) \quad (4)$$

with $l \in \mathbb{Z}$, the set of relative integers.

Now, considering that we are searching a probability distribution $\rho(x)$ and that M is the maximum of the genetic distance and that it is equal to 1, it is easy to see that $\rho(x)$ can be derived from $n(x)$ by setting l (in order to having no negative value for the distribution) to 1 and by computing B so that the probability distribution will be normalized. As a result, the distribution becomes:

$$\rho(x) = \frac{\pi}{2} \sin(\pi x) \quad (5)$$

And the probability of having a sequence with a genetic distance less or equal to x is given by the integration of (5):

$$P(X \leq x) = \frac{1}{2}[1 - \cos(\pi x)] \quad (6)$$

From genetic distance to sequence similarity

Unfortunately, we don't know the genetic distance between proteins (which is a difficult phylogenetic problem²⁰) but only the estimation of the Shared Amount of Information (SAI) between them. Usually, the SAI is given by a score which has been demonstrated to be an estimate of the mutual information I between the two biological sequence⁷ taken as events in a probability space and not as a random variable as in the theory of Shannon; In the discrete case, mutual information between random variables will be the average of mutual information between elementary events. To summarise, the SAI is the biological information shared by two proteins we want to access, the score is the mean to measure the SAI and the mutual information is the theoretical framework where the measure called "score" is a quantity properly defined. Finding a relationship between the genetic distance x and the mutual information I which conserves the essential properties can be

achieved by using an extremely simple model. First, we consider the variation of the genetic distance x with respect to the variation of I with the general model $dx/dI = f(x, I)$.

The variation of the genetic distance is clearly opposite to the variation of the mutual information between two sequences. The simplest model we could try is the model where the variation of the genetic distance is proportional to the variation of the mutual information and to the product of the genetic distance and the mutual information between the two biological sequences. With k a proportionality constant, we have:

$$dx = -kxIdI \quad (7)$$

which, after integration $\int_1^x \frac{dx}{x} = \int_{I(1)}^I -\kappa IdI$, lead to the following solution:

$$x(I) = \exp(-\alpha(I^2 - \xi^2)), I \in [\zeta, +\infty] \quad (8)$$

where $\alpha = \kappa/2$ and $\xi = I(1)$.

A new pairwise alignment score distribution

Applying the conservation of probability to the probability density given by equation (5), that is to say $\rho(x)dx = \rho(I)dI$ and so $\int_s^{+\infty} \rho(I)dI = P(S \geq s)$, we have $\int_0^x \rho(u)du = \int_s^{+\infty} \rho(I)dI$ which lead to the pairwise alignment score distribution, after replacing I by its estimate s :

$$\rho(s) = \frac{\alpha\pi}{2} s \exp(-\alpha(s^2 - \zeta^2)) \sin(\pi \exp(-\alpha(s^2 - \zeta^2))), \quad s \in [\zeta, +\infty] \quad (9)$$

which is strictly positive. Finally, we get the survival function $P(S \geq s) = 1/2[1 - \cos(\pi \exp(-\alpha(s^2 - \zeta^2)))]$ and so the repartition function

$$P(S \leq s) = \frac{1}{2}[1 + \cos(\pi \exp(-\alpha(s^2 - \zeta^2)))] \dots s \in [\zeta, +\infty], \quad (10)$$

which is a measure of probability since, with $F(s) = P(S \leq s)$, we have $F(\zeta) = 0$ and $F(+\infty) = 1$.

Comparisons between the new model and previous models

From the model (9) and (10), we can derive a more general family of density and probability distribution to be fitted to the data:

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

$$s \in [\zeta, +\infty] \quad (11)$$

$$P(S \leq s) = \frac{1}{2} [1 + \cos(\pi \exp(-\alpha(s^\eta - \zeta^\eta)))] , s \in [\zeta, +\infty] \quad (12)$$

To investigate the accuracy of this new model, we used two homologous Response Regulator NtrC family proteins in *Pseudomonas fluorescens* Pf-5 and in *Pseudomonas fluorescens* Pf0-1 (Accession numbers in the swissprot/UniProt database²⁶ PFL_0091 and Pfl01_0046). For a given number N of randomizations (here $N = 1000$), the second sequence Pfl01_0046 is uniformly shuffled (each sequence residue is permuted one time²⁷) and then aligned with the first sequence using the SIM algorithm²⁸ and the BLOSUM62 matrix⁵ with the NCBI Blast¹ default parameters (gap open penalty = 11, gap extension penalty = 1). This algorithm generates N random scores which will constitute the random score distribution to be compared with our model. Using uniform randomization is justified by the fact that adding neighbour constraints or using real database sequences (N database sequences chosen at random) leads generally to very similar results.^{9,27} We used three distributions to fit the random scores distribution: the classical Gumbel model $p(s) = 1/\beta \exp(-(s - \theta)/\beta) \cdot \exp(-\exp(-(s - \theta)/\beta))$ (which is equivalent to the model (1)), the gamma distribution model $p(s) = s^{\delta-1} \cdot \exp(-s/\omega) \cdot (\Gamma(\delta) \cdot \omega^\delta)^{-1}$ and our new model (11) with and with no preset for η . Indeed, the model (9) assigned a value of 2 for this parameter. As a consequence, we hope that the value of η optimizing from the data will be close to 2. Optimisation of the parameters θ and β for the Gumbel model (1); δ and ω for the gamma distribution model; α , ζ and η for the our model was achieved using maximum likelihood statistical inference and was implemented in the R statistical language.²⁹ The maximum likelihood approach has

been chosen because it can be easily written and implemented.³⁰ All optimisation was done using the nlm package available with the R statistical software.²⁹ For this work, an R library, called *basto.r*, has been developed. It includes an implementation of the Heaviside function and a set of functions for the density, the cumulative function, the generator of random values and the log-likelihood for the new model (the R library corresponding to the new model (11) is available in the supplementary materials and is part of the *basto.r* package's *bastog* family functions). As it can be seen in Figure 2, the new model qualitatively outperforms the Gumbel model. This can be verifying in Figure 3 by using a qq-plot representation. In addition, the new model is the only one which fits data in the twilight zone, which is the zone of low protein alignment scores. The model (11) has three parameters whereas the Gumbel model (1) and the gamma model have both two parameters. Classically, it isn't surprising that the model (11) could have a better fit with the data while possessing a greater number of parameters than the other models. However, we can observe that the optimized

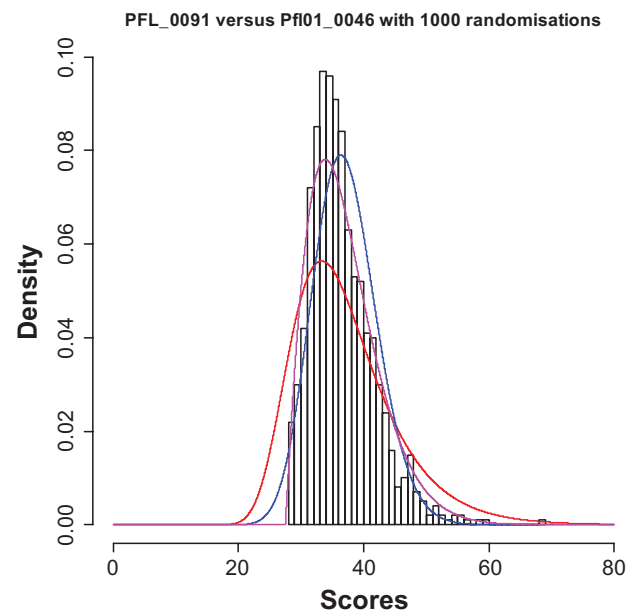


Figure 2. Histogram of the distribution of random scores between the Response Regulator NtrC family proteins in *Pseudomonas fluorescens* Pf-5 and homologous proteins in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_0091 and Pfl01_0046). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 33.27876$ and $\beta = 6.523116$. Blue curve: gamma distribution with parameters $\delta = 53.04861$ and $\omega = 0.6983029$. Purple curve: our model with parameters $\alpha = 0.001281424$, $\zeta = 27.500000245$ and $\eta = 1.999992822$.

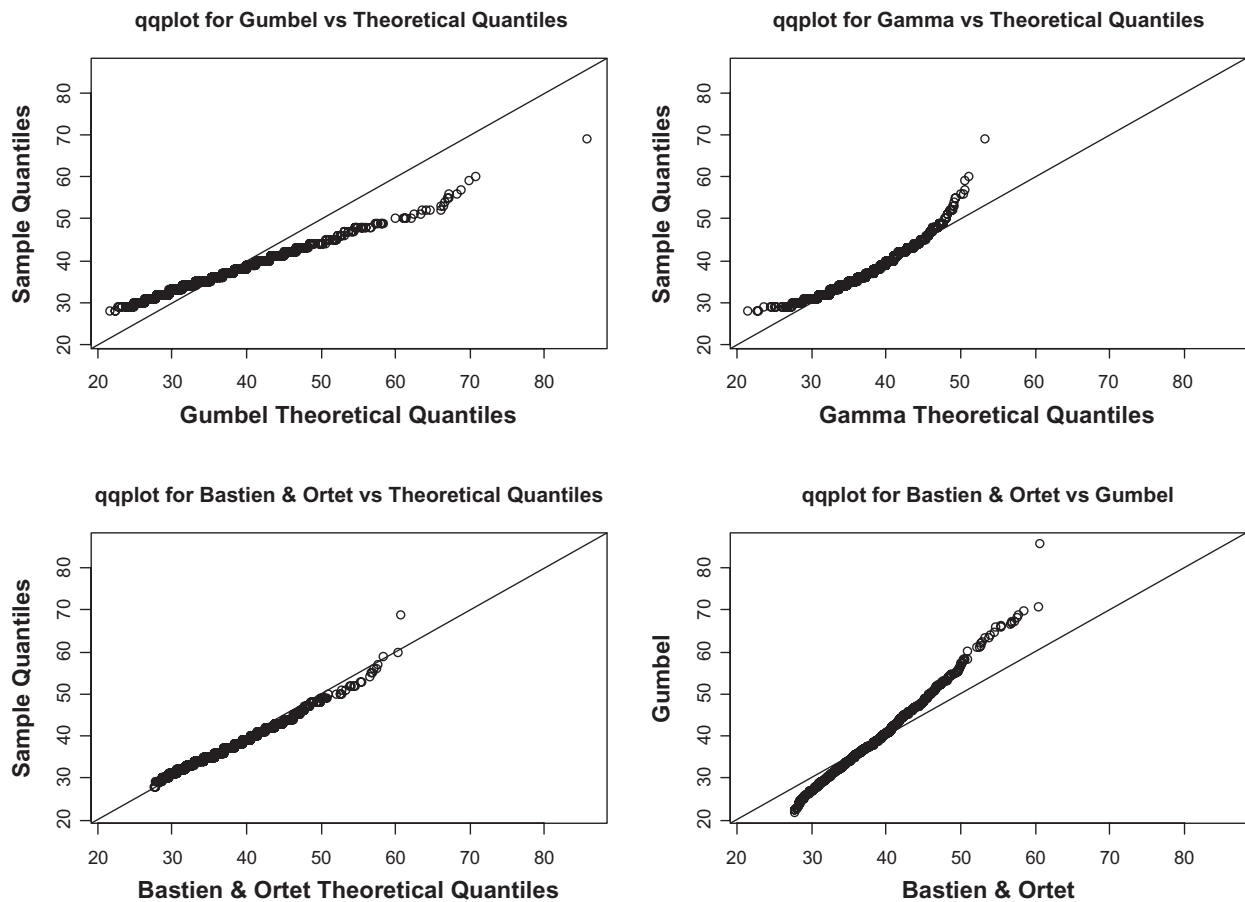


Figure 3. Quantile-quantile plot of (**top-left**) the Gumbel theoretical quantile versus the data quantile. (**top-right**) the gamma theoretical quantile versus the data quantile. (**down-left**) our model theoretical quantile versus the data quantile. (**top-left**) our model theoretical quantile versus the Gumbel theoretical quantile.

value of $\eta = 1.999992822$ is very close from that of the model (9). As a consequence, the two parameter model (9) seems also to be a better model than the Gumbel and the gamma model. This assumption is confirmed by the results of parameter optimizations (supplementary materials, p. 6) with the constraint $\eta = 2$ (basto.r package's basto2 family functions).

A very interesting feature of this new model is that it combines the advantages of both the Gumbel law and the gamma law. Indeed, The Karlin-Altschul domain of validity is precisely the domain of sufficiently high score (and/or long length sequences as can be seen in equation(1)).^{10,17} In fact, the Gumbel model is less efficient than the gamma or the normal distribution in the low score domain¹⁵ for global alignment scores but it is more efficient in the high score domain.^{10,12,17} The new proposed model seems to have properties to reflect the distribution of random scores both in the low score region (the so-called twilight zone) and in the high score region, the

common regions of interest. Three more examples are given by the pairwise alignments of three proteins in *Pseudomonas fluorescens* Pf-5 and in *Pseudomonas fluorescens* Pf0-1: a two-component system protein from the NarL family (Fig. 4) ($\eta = 2.053335$), the rod shape-determining protein MreB (Fig. 5) ($\eta = 1.908571$) and a hypothetical protein chosen because it possesses a particularly high score with its homologue (Fig. 6) ($\eta = 2.5$). However, this conclusion applies to all existing pairwise alignments score distributions.

Discussion

This model provides the first link between protein sequence comparison results and the evolutionary processes which have led to those proteins. Starting from a simple evolutionary process, we obtained a score probability distribution (9) and its generalized version (11). This new model seems more accurate than other previously tested models. However, the determination of the

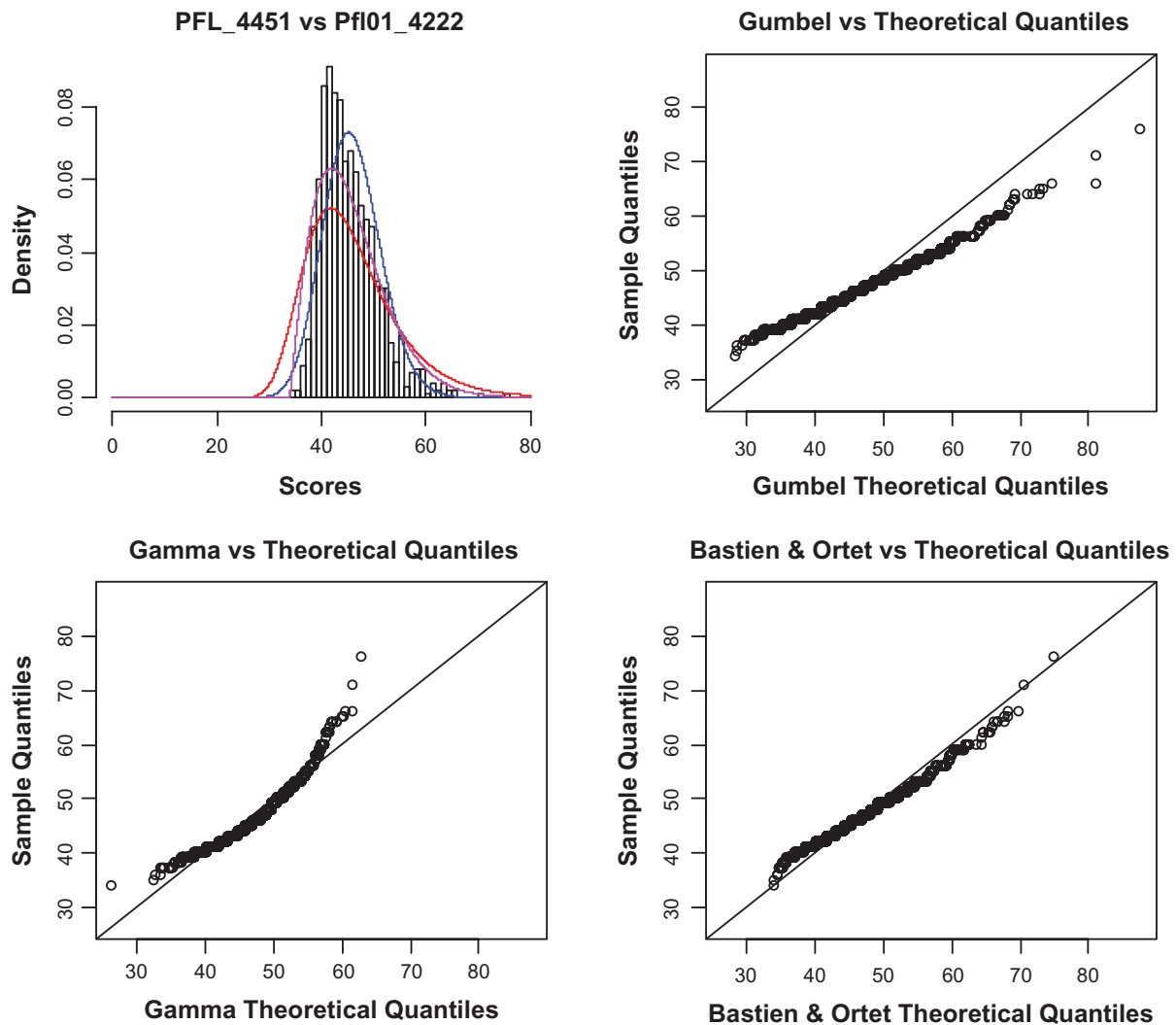


Figure 4. (Top Left) Histogram of the distribution of random scores between the two-component system, NarL family, sensor histidine kinase in *Pseudomonas fluorescens* Pf-5 and homologous proteins in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_4451 and Pfl01_4222). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 41.79989$ and $\beta = 7.047815$. Blue curve: gamma distribution with parameters $\delta = 69.67213$ and $\omega = 0.6583407$. Purple curve: our model with parameters $\alpha = 0.0006687308$, $\zeta = 33.99296$ and $\eta = 2.053335$. **(top-right)** Quantile-quantile plot of the gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.

parameters of the Gumbel distribution is a computationally expensive task^{9,17,31,32} although several efforts have reduced this expense by algorithm improvements,^{31–33} or new sample statistical procedures.³⁴ The statistical estimation of the distribution also varies with the chosen substitution matrix and the chosen alignment algorithm.^{5,6,35–37} As a consequence, further work on the new models (9) and (11) should include the influence of the substitution matrices and the construction of an efficient algorithm to determine the models' parameters. Indeed, the application to a database search using maximum likelihood methods is unrealistic because of the extensive time required.

In order to state the domain of validity of this new model, it will be important to investigate the influence of allowing τ to vary over time in the divergence-duplication model. Indeed, the evolutionary model used in this work is a branching process which has to be related with both graph theory and graphical, often tree, representation of evolution. As a consequence, we hope that suspending the hypothesis that τ is constant will lead to a more realistic and accurate model. A second point which should be studying in depth is the influence of the sequence alignment parameter-settings. Nevertheless, preliminary studies tend to show

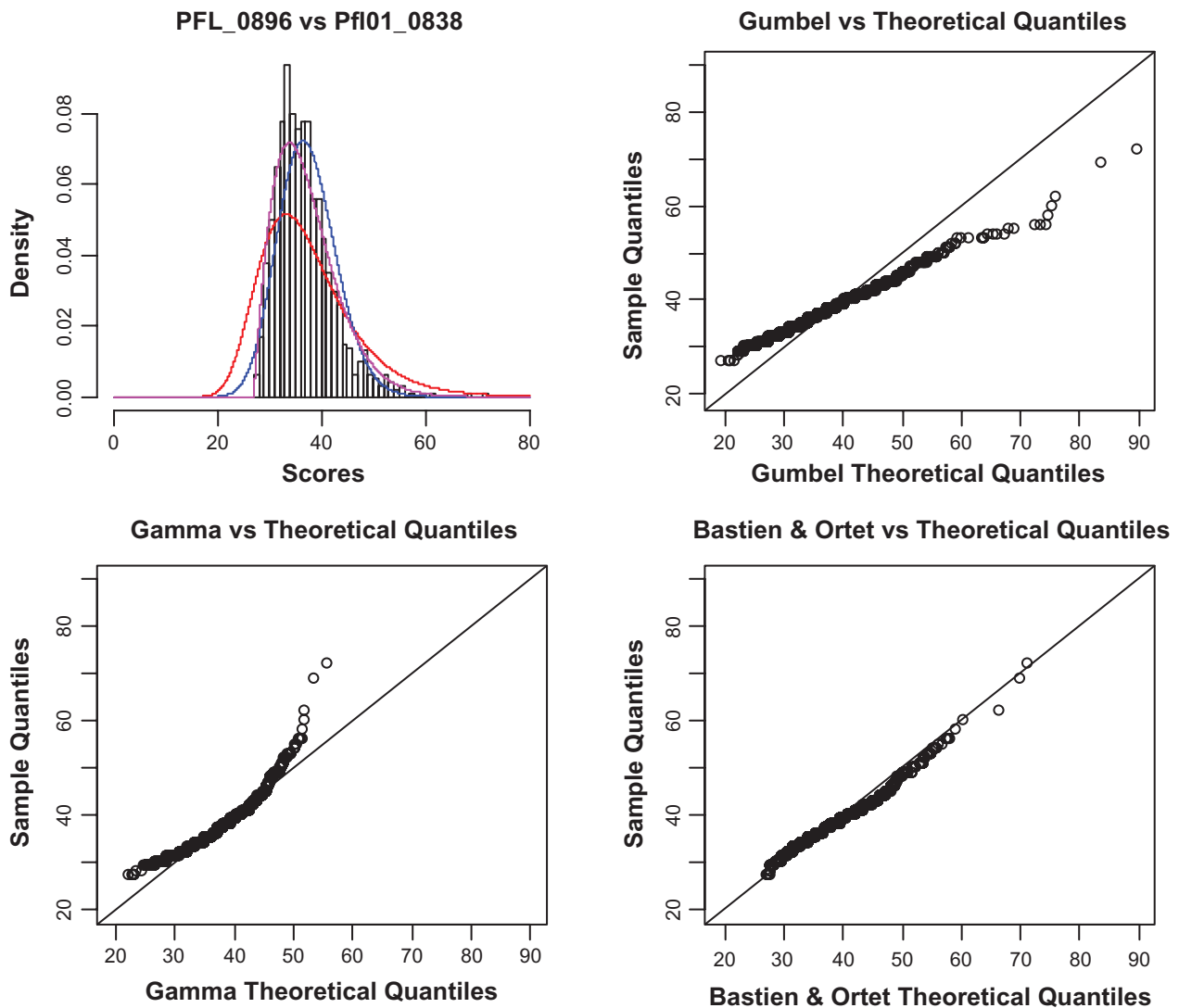


Figure 5. (Top Left) Histogram of the distribution of random scores between the rod shape-determining protein MreB in *Pseudomonas fluorescens* PF-5 and the homologous protein in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_0896 and Pfl01_0838). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 33.20788$ and $\beta = 7.1206$. Blue curve: gamma distribution with parameters $\delta = 45.1806$ and $\omega = 0.825974$. Purple curve: our model with parameters $\alpha = 0.001707993$, $\zeta = 26.95872$ and $\eta = 1.908571$. **(top-right)** Quantile-quantile plot of the Gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.

that the above conclusions hold for comparisons of sequences of various lengths and also for different parameters-setting (supplementary materials, p. 11). It will also be interesting to test different relations between mutual information and genetic distance (equation 7)³⁸ to evaluate the robustness of the model. In this work, we tried to retrieve the Gumbel distribution shape from a purely evolutionary process and we obtained a new score probability distribution which exhibits great statistical accuracy. However, many bioinformatics applications are also interesting in separating related pairs

of sequences from unrelated ones, like in a database search. A retrieval accuracy study should also be undertaken in future work. An investigation of the possible extension of this distribution to global alignments should also be undertaken.

Acknowledgements

The authors thank Dr. David Whitworth for proof-reading the manuscript for grammar and spelling. This work was supported by grants from the Agence Nationale de la Recherche, as part of the PlasmExplore project.

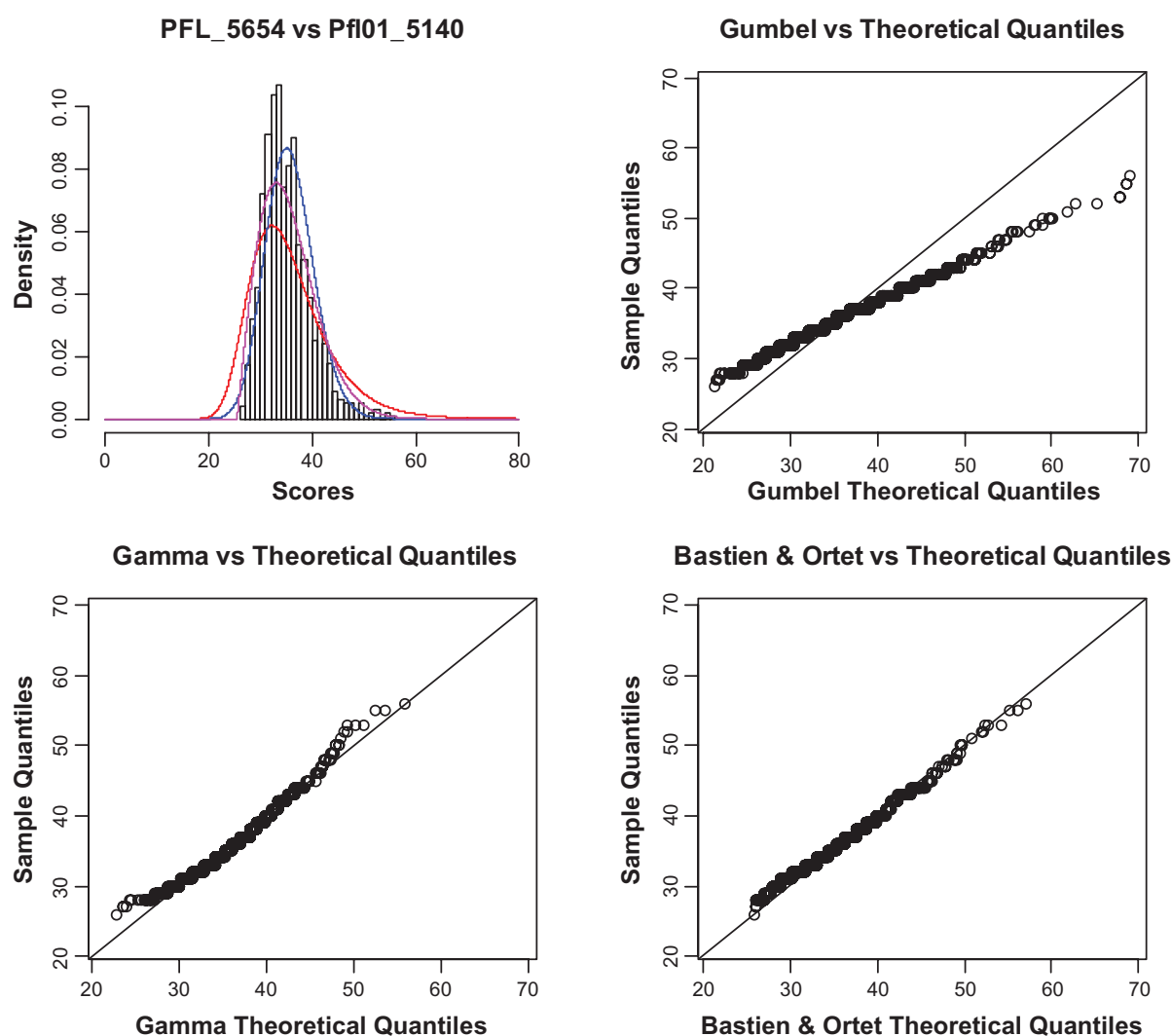


Figure 6. (Top Left) Histogram of the distribution of random scores between a hypothetical protein in *Pseudomonas fluorescens* Pf-5 and the homologous protein of *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_5654 and Pf01_5140). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 32.25266$ and $\beta = 5.941174$. Blue curve: gamma distribution with parameters $\delta = 59.33383$ and $\omega = 0.601377$. Purple curve: our model with parameters $\alpha = 0.0001799611$, $\zeta = 25.5$ and $\eta = 2.5$. **(top-right)** Quantile-quantile plot of the Gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- Altschul SF, Madden TL, Schaer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acid Res.* 1997;25(17):3389.
- Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147:195–7.
- Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. *Atlas Prot Seq Struct.* 1978;5:345–52.
- Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 1992;89:10915–9.
- Bastien O, Roy S, Marechal E. Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C R Biol.* 2005;328(5):445–53.
- Bastien O, Ortet P, Roy S, Maréchal E. A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. *BMC Bioinformatics.* 2005;6:49.
- Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 1991;219(3):555–65.
- Mitrophanov AY, Borodovsky M. Statistical significance in biological sequence analysis. *Brief Bioinformatics.* 2006;7(1):2–24.
- Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A.* 1990;87:2264–8.



11. Altschul SF, Bundschuh R, Olsen R, Hwa T. The estimation of statistical parameters for local alignment score distributions. *Nucl Acid Res.* 2001;29:351–61.
12. Pearson WR. Empirical statistical estimates for sequence similarity searches. *J Mol Biol.* 1998;276:71–84.
13. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol.* 2008;4(5):e1000069.
14. Poleksic A. Island method for estimating the statistical significance of profile-profile alignment scores. *BMC Bioinformatics.* 2009;10:112.
15. Pang H, Tang J, Chen SS, Tao S. Statistical distributions of optimal global alignment scores of random protein sequences. *BMC Bioinformatics.* 2005;6:257.
16. Webber C, Barton GJ. Estimation of *P*-values for global alignments of protein sequences. *Bioinformatics.* 2001;17:1158–67.
17. Comet JP, Aude JC, Glemet E, et al. Significance of Z-value statistics of Smith—Waterman scores for protein alignments. *Comput Chem.* 1999;23:317–31.
18. Bastien O. A Simple Derivation of the Distribution of Pairwise Local Protein Sequence Alignment Scores. *Evolutionary Bioinformatics.* 2008;4:41.
19. Bastien O, Maréchal E. Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores. *BMC Bioinformatics.* 2008;9:332.
20. Salemi M, Vandamme AM. *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.* Cambridge: Cambridge University Press; 2003.
21. <http://mathworld.wolfram.com/PolyasRandomWalkConstants.html>
22. Finch SR. *Mathematical Constants.* Cambridge: Cambridge University Press; 2003.
23. Alonso M, Finn EJ. *Physics.* Reading Massachusetts: Addison-Wesley; 1992.
24. Cohen-Tannoudji C, Diu B, Laloë F. *Mécanique quantique I.* Paris: Hermann; 1997.
25. Teschl G. *Quantum Mechanics and Applications to Schrödinger Operators.* Rhode Island: American Mathematical Society; 2009.
26. <http://expasy.org/sprot/>
27. Fitch WM. Random sequences. *J Mol Biol.* 1983;163(2):171–6.
28. Huang X, Miller W. A Time-Efficient, Linear-Space Local Similarity Algorithm. *Advances in Applied Mathematics.* 1991;12:337–57.
29. Ihaka R, Gentleman RR. A language for data analysis and graphics. *J Comput Graph Stat.* 1996;5:299–314.
30. Fourdrinier D. *Statistique Inférentielle.* Paris: Dunod; 2002.
31. Waterman MS, Vingron M. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc Natl Acad Sci U S A.* 1994;91(11):4625.
32. Bundschuh R. Rapid significance estimation in local sequence alignment with gaps. *J Comput Biol.* 2002;9(2):243–60.
33. Chia N, Bundschuh R. *Research in Computational Molecular Biology.* Berlin: Springer; 2005.
34. Newberg LA. Significance of gapped sequence alignments. *J Comput Biol.* 2008;15(9):1187–94.
35. Schaffer AA, Aravind L, Madden TL, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl Acid Res.* 2001;29(14):2994–3005.
36. Agrawal A, Brendel VP, Huang X. Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment. *Int J Comput Biol Drug Design.* 2008;1(4):347–67.
37. Agrawal A, Huang X. Pairwise statistical significance of local sequence alignment using multiple parameter sets and empirical justification of parameter set change penalty. *BMC Bioinformatics.* 2009;10(3):S1.
38. Grishin NV. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J Mol Evol* 1995;41(5):675–9.



Supplementary Materials

Table of Contents

The basto R library (page 2–4) must be copying in a file named basto.r (ascii format)	2
Example of the basto.r usage; procedure	5
Example of the basto.r usage; source code	5
Parameter Optimizations with $\eta = 2$	6
Alignment Parameter Influence Study	10



```
#basto library
#Author: Dr Olivier Bastien
#email: olivier.bastien@cea.fr
#date: 25 April 2010
#
#must be copying in a file named basto.r (ascii format)
#heaviside function
#x: function support
#y: function values
#cutoff: the heaviside value of the function vanishes below this cutoff
heaviside<-function(x,y,cutoff = 0)
{
Z<-0;
sizeX<-length(x);
for(i in 1:sizeX)
{
if(x[i]< cutoff)
{
Z[i]<-0;
}
else
{
Z[i]<-y[i];
}
}
Z
}

#basto distribution
#Density, distribution function, quantile function and random generation
#for the basto distribution
#x, q: vector of quantiles
#p: vector of probabilities
#n: number of observations
#alpha: vector of alpha
#xmin: vector of xmin
dbasto<-function(x,alpha = 0.2,xmin = 0)
{
heaviside(x, (alpha*pi/2)*exp(-alpha*(x-xmin))*sin(pi*exp(-alpha*(x-xmin))),xmin)
}
pbasto<- function(q,alpha = 0.2,xmin = 0)
{
heaviside(q, (1/2)*(1+cos(pi*exp(-alpha*(q-xmin))))),0)
}
qbasto<-function(p,alpha = 0.2,xmin = 0)
```



```

{
}
rbasto<-function(n,alpha = 0.2,xmin = 0)
{
Z<-runif(n,min = 0,max = 1);
-(1/alpha)*log((1/pi)*acos((2*Z)-1))+xmin
}
#log maximum likelihood for the basto distribution
# vector of parameters which is c(alpha,xmin)
mlbasto<-function(para,TT)
{
alpha<-para[1];
xmin<-para[2];
LL<-length(TT);
-sum(log(dbasto(TT,alpha,xmin)))
}

#generalized basto distribution
#Density, distribution function, quantile function and random generation
#for the generalized basto distribution
#x, q: vector of quantiles
#p: vector of probabilities
#n: number of observations
#alpha: vector of alpha
#xmin: vector of xmin
dbastog<-function(x,alpha = 0.2,xmin = 0,beta = 1)
{
#heaviside(x, (alpha*pi/2)*exp(-alpha*(x-xmin))*sin(pi*exp(-alpha*
(x-xmin))),xmin)
heaviside(x, (beta*alpha*pi/2)*(x^(beta-1))*exp(-alpha*(x^(beta)-
xmin^(beta)))*sin(pi*exp(-alpha*(x^(beta)-xmin^(beta)))),xmin)
}
rbastog<- function(n,alpha = 0.2,xmin = 0,beta = 1)
{
Z<-runif(n,min = 0,max = 1);
(-(1/alpha)*log((1/pi)*acos((2*Z)-1))+xmin^(beta))^(1/beta)
}
#log maximum likelihood for the generalized basto distribution
#vector of parameters which is c(alpha,xmin)
mlbastog<-function(para,TT)
{
alpha<-para[1];
xmin<-para[2];
beta<-para[3];
LL<-length(TT);
-sum(log(dbastog(TT,alpha,xmin,beta)))
}

```



```
#basto2 distribution
#Density, distribution function, quantile function and random generation
#for the generalized basto distribution
#x, q: vector of quantiles
#p: vector of probabilities
#n: number of observations
#alpha: vector of alpha
#xmin: vector of xmin
dbasto2<-function(x,alpha = 0.2,xmin = 0)
{
#heaviside(x, (alpha*pi/2)*exp(-alpha*(x-xmin))*sin(pi*exp(-alpha*
(x-xmin))),xmin)
heaviside(x, (2*alpha*pi/2)*(x^(2-1))*exp(-alpha*(x^(2)-
xmin^(2)))*sin(pi*exp(-alpha*(x^(2)-xmin^(2)))),xmin)
}
rbasto2<-function(n,alpha = 0.2,xmin = 0)
{
Z<-runif(n,min = 0,max = 1);
(-(1/alpha)*log((1/pi)*acos((2*Z)-1))+xmin^(2))^(1/2)
}
#log maximum likelihood for the generalized basto distribution
#vector of parameters which is c(alpha,xmin)
mlbasto2<-function(para,TT)
{
alpha<-para[1];
xmin<-para[2];
LL<-length(TT);
-sum(log(dbasto2(TT,alpha,xmin)))
}
```



Example of the basto.r usage (page 5)

1. Copy the example source code (below) in a file named test_basto.r (ascii format).
2. Verify that the two files test_basto.r and the basto.r library, are in the same repertory.
3. Just after, launch the R software and write the command line: >source ("test_basti.r")

```
source("basto.r")
XX<-rbastog(1000,0.002,20,2)
hist(XX,nclass = 50,,xlim = c(0,100),freq = FALSE);
test.optim<-nlm(mlbastog,c(0.0001,min(XX)-1,2),XX,print.level = 2);
X<-seq(from = 0,to = 80,by = 0.01);
funcTest<-dbastog(X,test.optim$estimate[1],test.optim$estimate[2],test.
optim$estimate[3]);
lines(X,funcTest,type = "l",col = "blue")
X11()
YY<-rbastog(100, test.optim$estimate[1],test.optim$estimate[2],test.
optim$estimate[3]);
minG<-min(c(XX,YY));
maxG<-max(c(XX,YY));
qqplot(XX,YY,xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
xlim = c(minG,maxG),ylim = c(minG,maxG))
```


Parameters Optimizations with $\eta = 2$

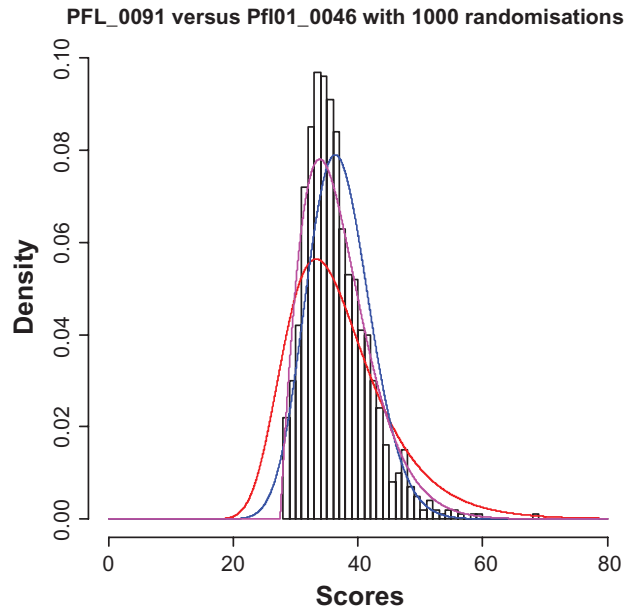


Figure S1. Histogram of the distribution of random scores between the Response Regulator NtrC family proteins in *Pseudomonas fluorescens* Pf-5 and homologous proteins in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_0091 and Pf01_0046). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 33.27876$ and $\beta = 6.523116$. Blue curve: gamma distribution with parameters $\delta = 53.04861$ and $\omega = 0.6983029$. Purple curve: our model with parameters $\alpha = 0.001281286$, $\zeta = 27.500000244$ and $\eta = 2$.

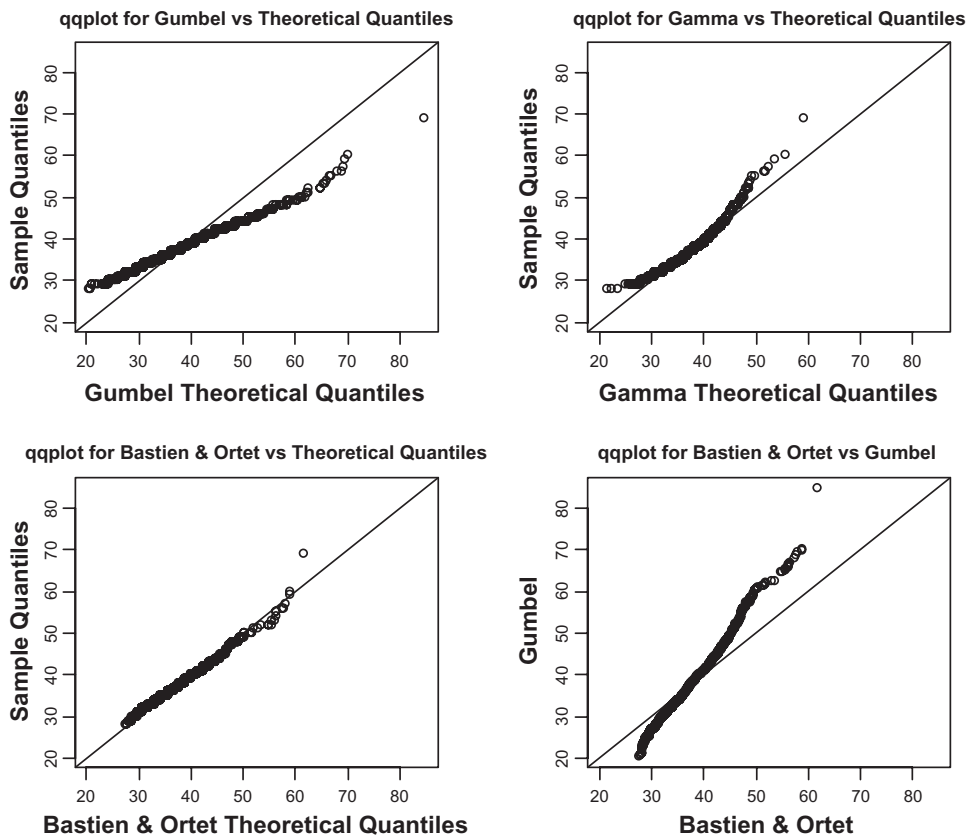


Figure S2. Quantile-quantile plot of (top-left) the Gumbel theoretical quantile versus the data quantile. (top-right) the gamma theoretical quantile versus the data quantile. (down-left) our model theoretical quantile versus the data quantile. (bottom-right) our model theoretical quantile versus the Gumbel theoretical quantile.

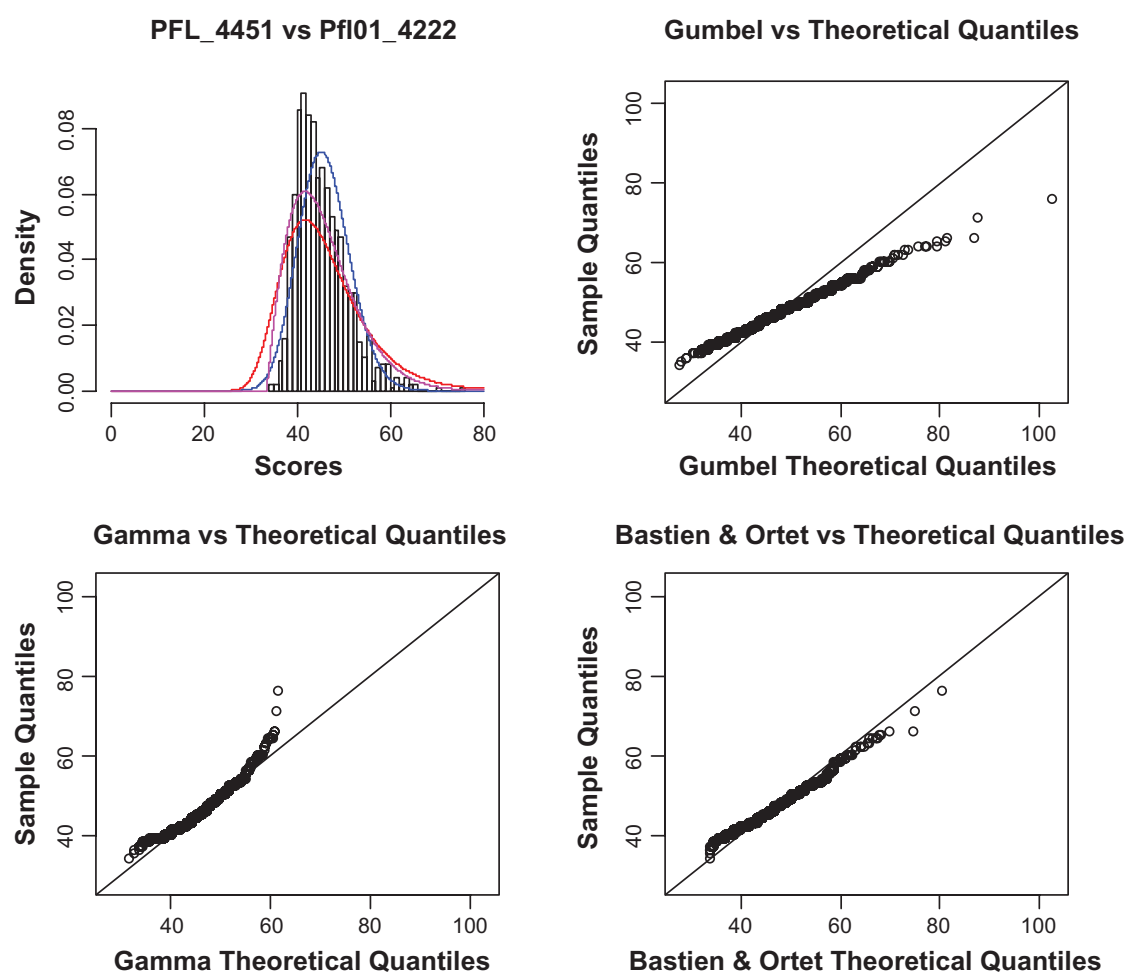


Figure S3. (Top Left) Histogram of the distribution of random scores between the two-component system, NarL family, sensor histidine kinase in *Pseudomonas fluorescens* Pf-5 and homologous proteins in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_4451 and Pf101_4222). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 41.79989$ and $\beta = 7.047815$. Blue curve: gamma distribution with parameters $\delta = 69.67213$ and $\omega = 0.6583407$. Purple curve: our model with parameters $\alpha = 8.094621e-04$, $\zeta = 33.5$ and $\eta = 2$. **(top-right)** Quantile-quantile plot of the gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.

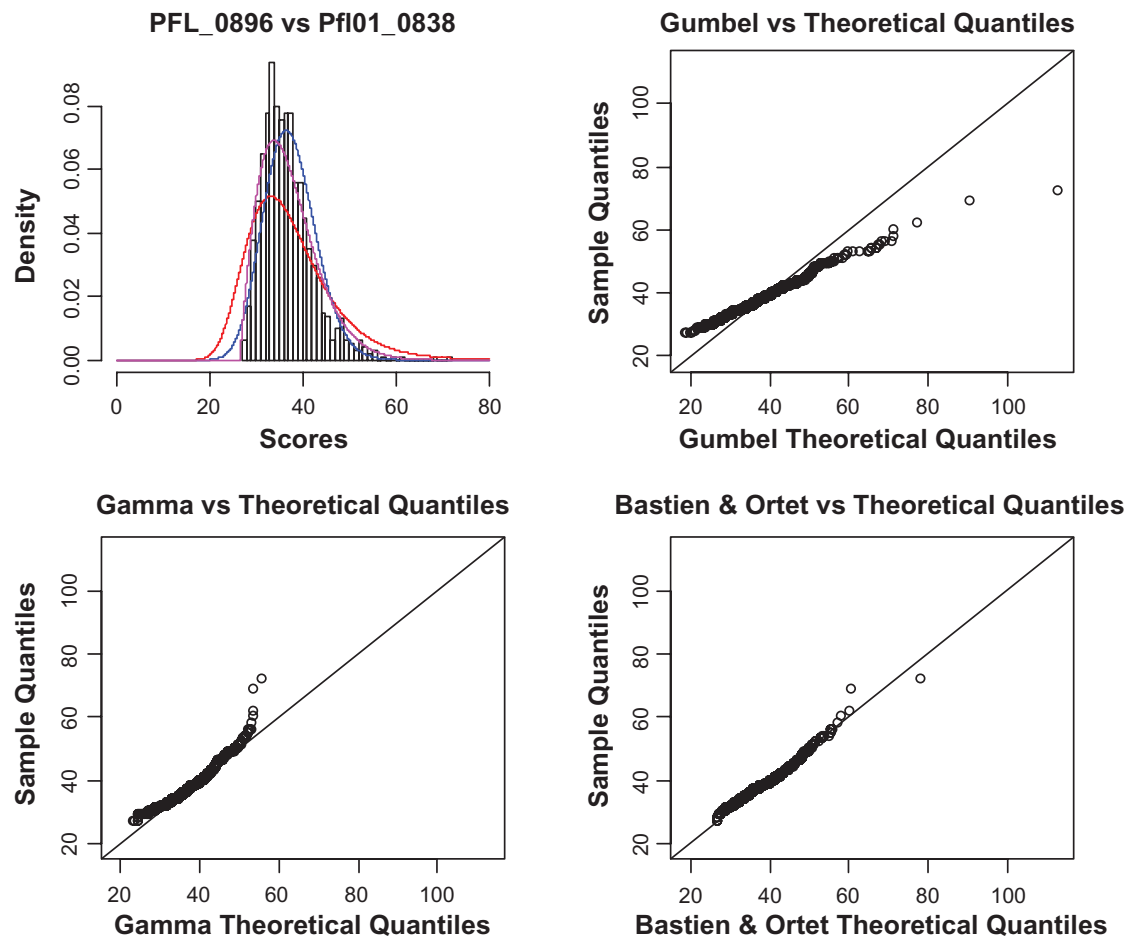


Figure S4. (Top Left) Histogram of the distribution of random scores between the rod shape-determining protein MreB in *Pseudomonas fluorescens* PF-5 and the homologous protein in *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_0896 and Pfl01_0838). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 33.20788$ and $\beta = 7.1206$. Blue curve: gamma distribution with parameters $\delta = 45.1806$ and $\omega = 0.825974$. Purple curve: our model with parameters $\alpha = 0.001140583$, $\zeta = 26.5$ and $\eta = 2$. **(top-right)** Quantile-quantile plot of the Gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.

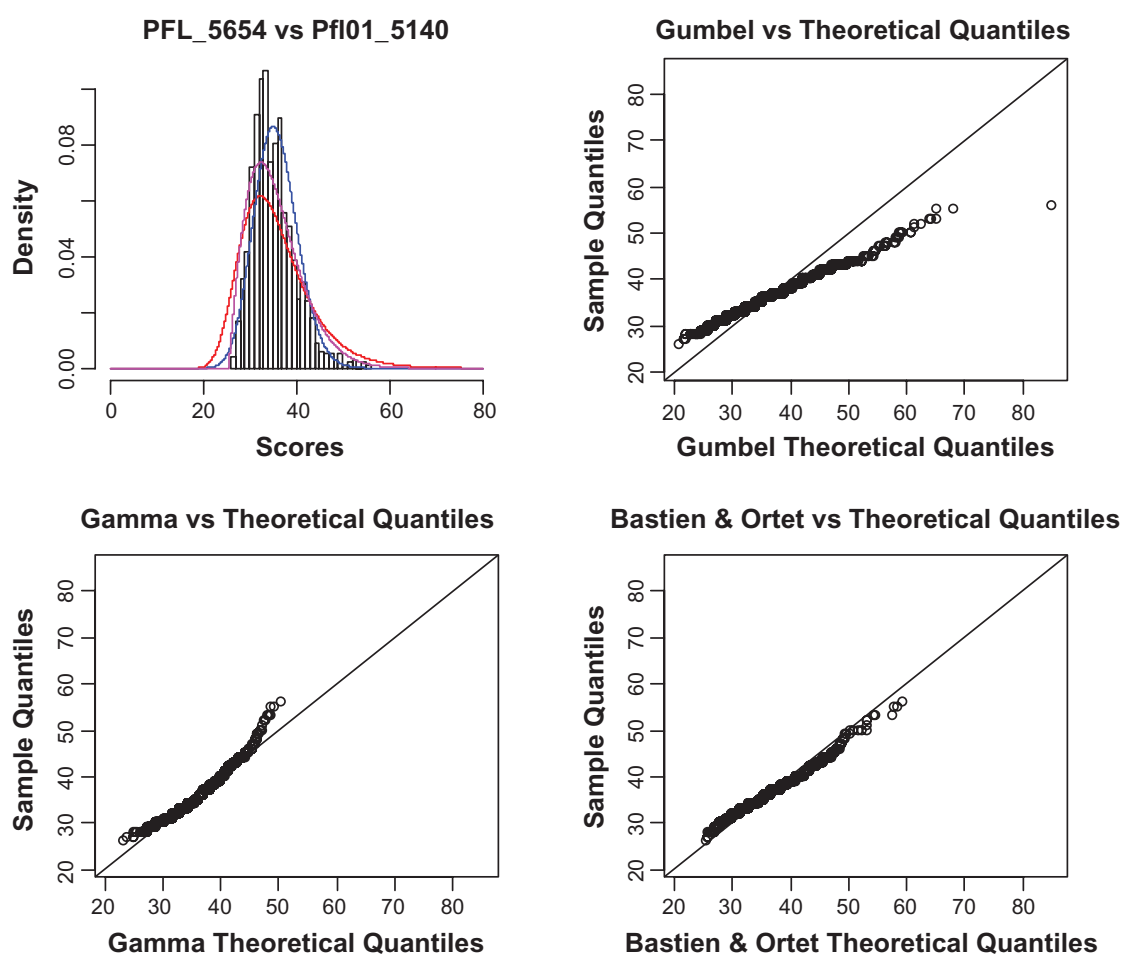


Figure S5. (Top Left) Histogram of the distribution of random scores between a hypothetical protein in *Pseudomonas fluorescens* Pf-5 and the homologous protein of *Pseudomonas fluorescens* Pf0-1 (Accession numbers PFL_5654 and Pf01_5140). Only the second sequence was shuffled 1000 times. Red curve: Gumbel distribution with parameters $\theta = 32.25266$ and $\beta = 5.941174$. Blue curve: gamma distribution with parameters $\delta = 59.33383$ and $\omega = 0.601377$. Purple curve: our model with parameters $\alpha = 0.001270900$, $\zeta = 25.5$ and $\eta = 2$. **(top-right)** Quantile-quantile plot of the Gumbel theoretical quantile versus the data quantile. **(down-left)** the gamma theoretical quantile versus the data quantile. **(down-right)** our model theoretical quantile versus the data quantile.



Alignment Parameter Influence Study

Comparison of a *Pseudomonas fluorescens* Pf-5 protein with four others *Pseudomonas fluorescens* Pf0-1 proteins

1. Liste of proteins

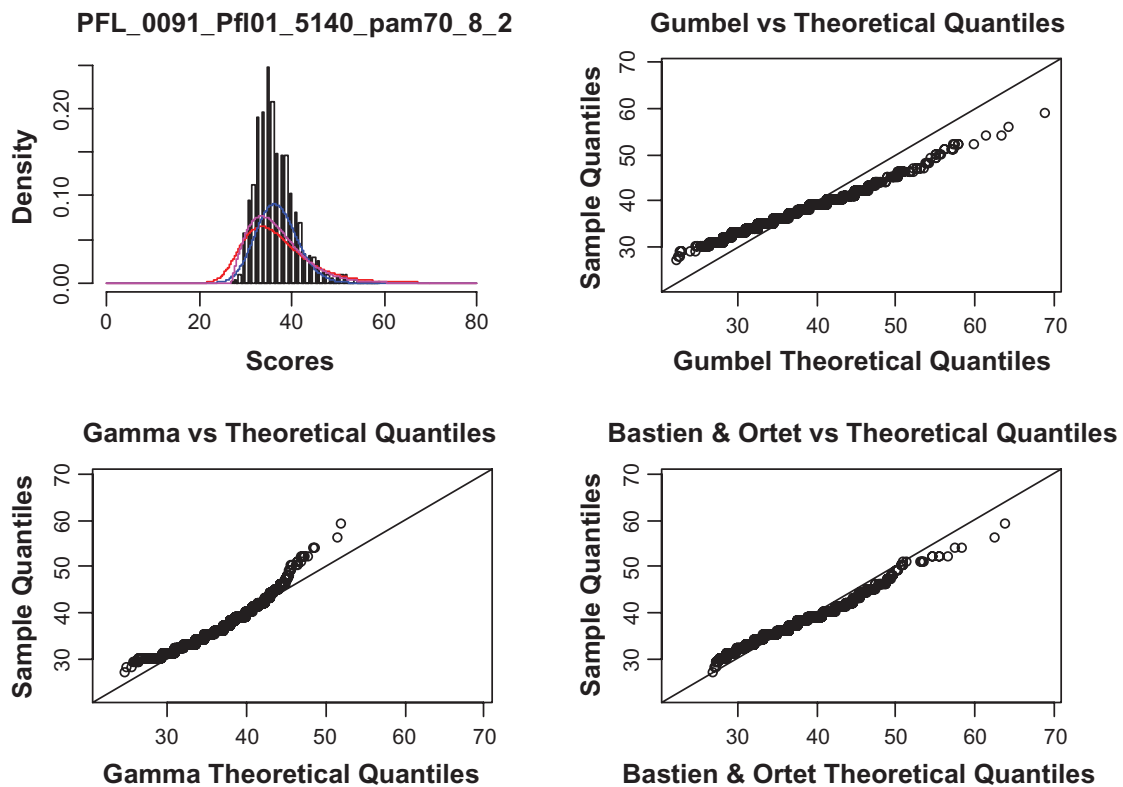
Pseudomonas fluorescens Pf-5	PFL_0091	448 aa
Vs Pseudomonas fluorescens Pf0-1	PfI01_0046	448 aa
Pseudomonas fluorescens Pf0-1	PfI01_4222	922 aa
Pseudomonas fluorescens Pf0-1	PfI01_0838	345 aa
Pseudomonas fluorescens Pf0-1	PfI01_5140	423 aa

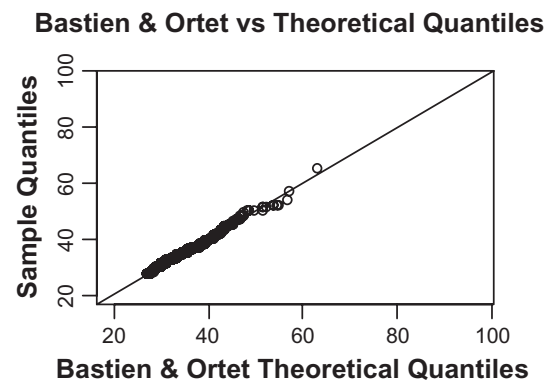
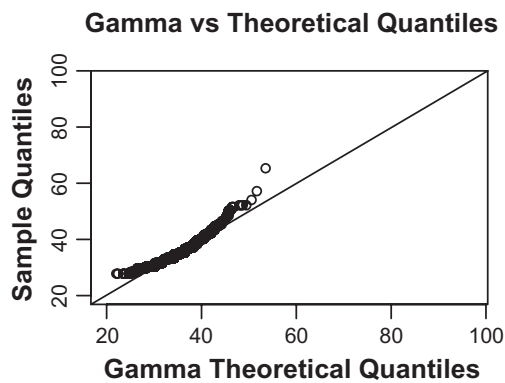
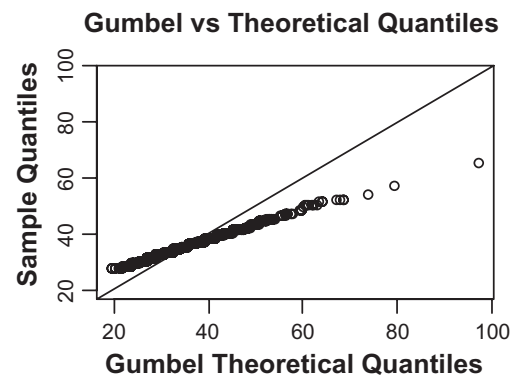
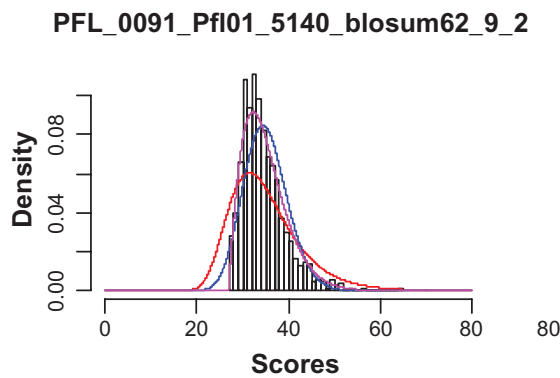
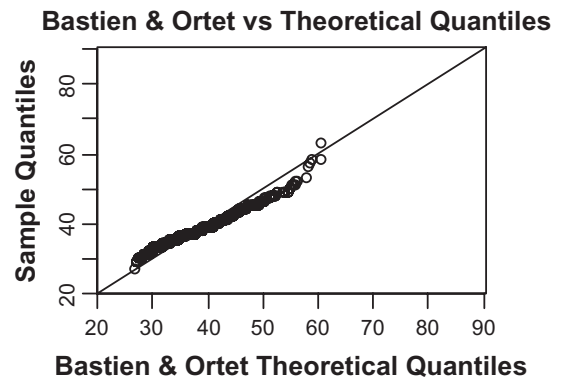
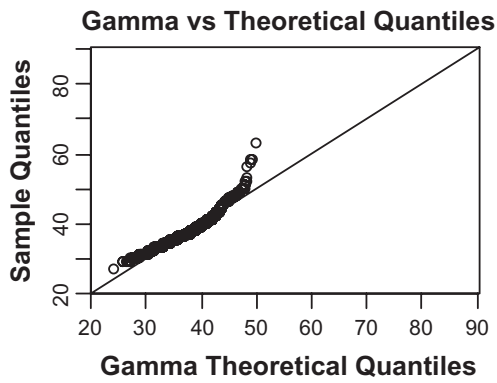
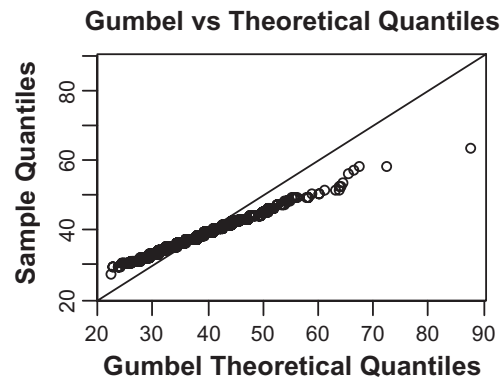
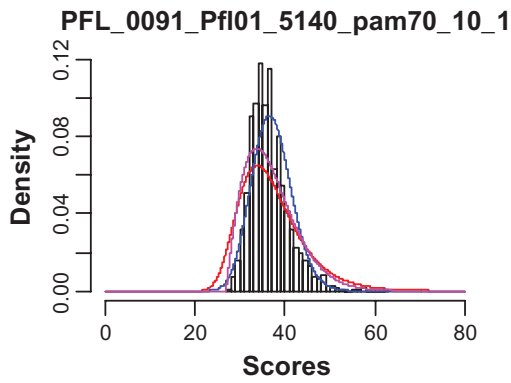
2. List of substitution matrices and alignment parameters (Gapo: gap open penalty, Gape: gap extension penalty)

Substitution matrices	First choice of parameters	Second choice of parameters
Blosum 62	Gapo: 11; Gape: 1	Gapo: 9; Gape: 2
Pam 70	Gapo: 10; Gape: 1	Gapo: 8; Gape: 2

3. Results

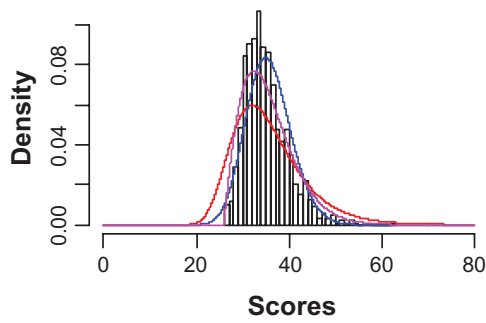
Title of figures: ID QUERY_ID SUBJET_ MATRICE_ GAPO_ GAPE



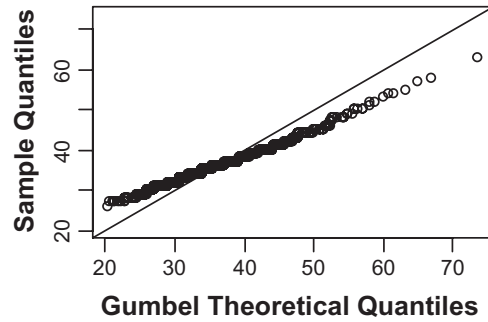




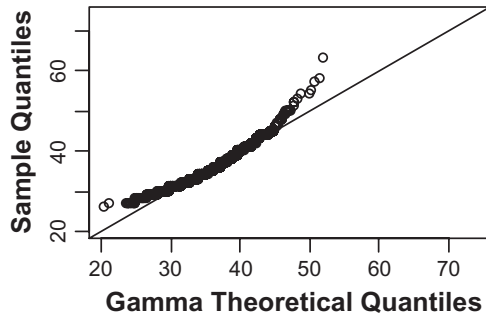
PFL_0091_PfI01_5140_blosum62_11_1



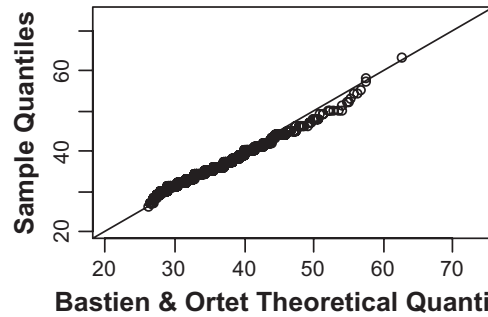
Gumbel vs Theoretical Quantiles



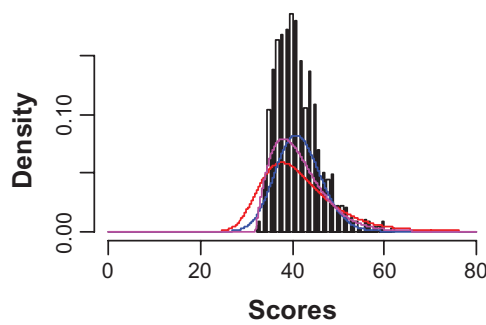
Gamma vs Theoretical Quantiles



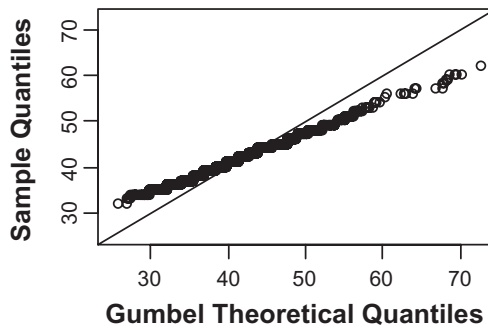
Bastien & Ortet vs Theoretical Quantiles



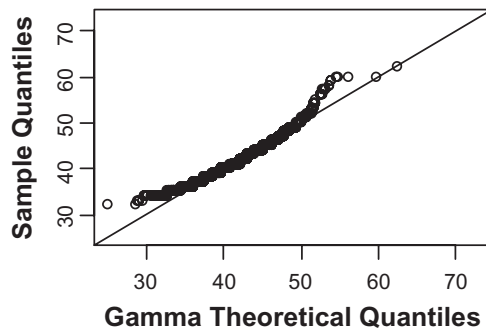
PFL_0091_PfI01_4222_pam70_8_2



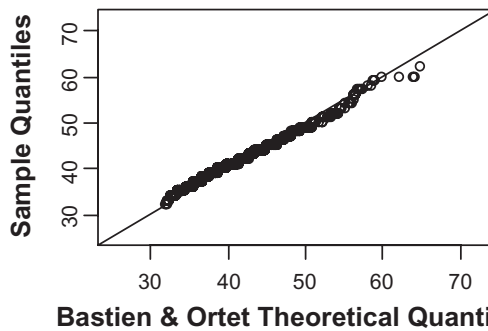
Gumbel vs Theoretical Quantiles



Gamma vs Theoretical Quantiles

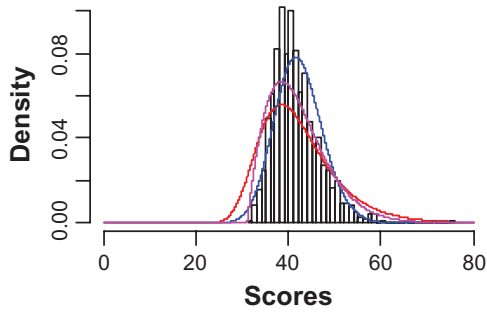


Bastien & Ortet vs Theoretical Quantiles

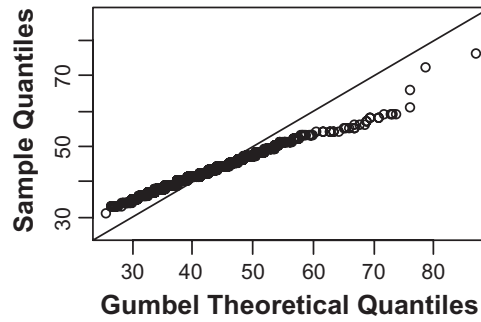




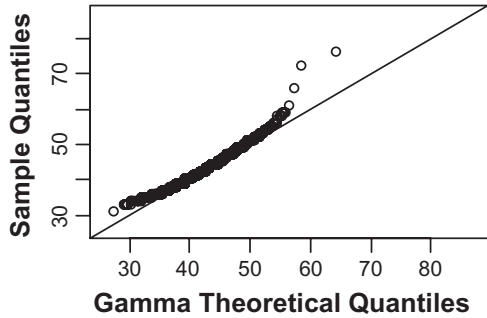
PFL_0091_Pf101_4222_pam70_10_1



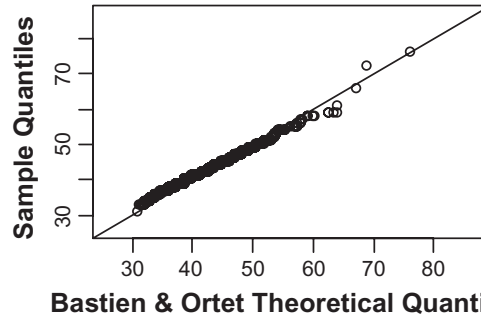
Gumbel vs Theoretical Quantiles



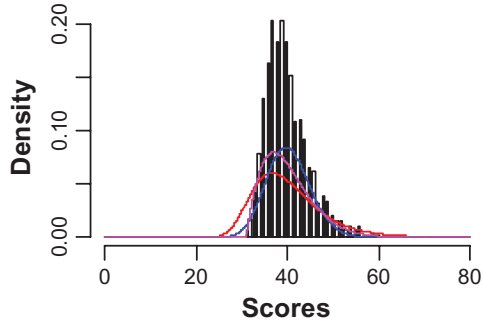
Gamma vs Theoretical Quantiles



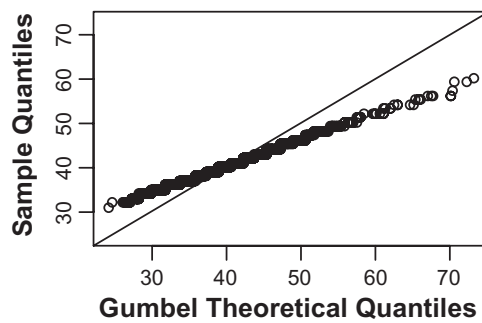
Bastien & Ortet vs Theoretical Quantiles



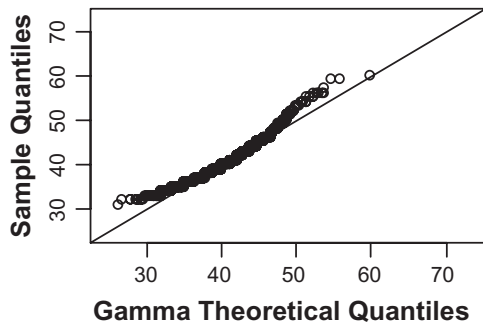
PFL_0091_Pf101_4222_blosum62_9_2



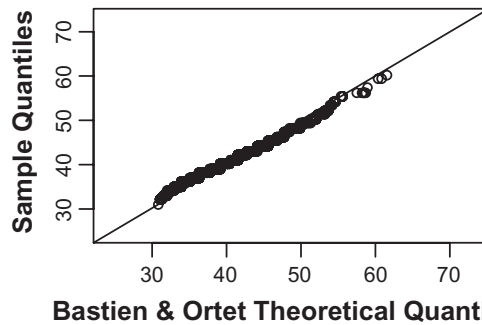
Gumbel vs Theoretical Quantiles

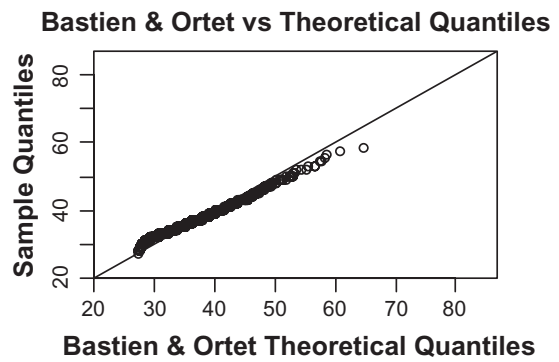
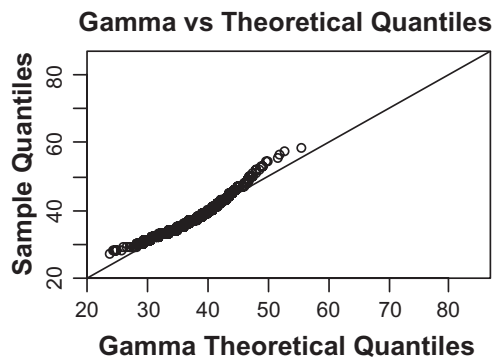
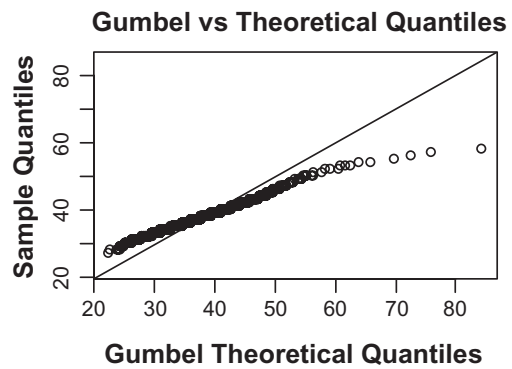
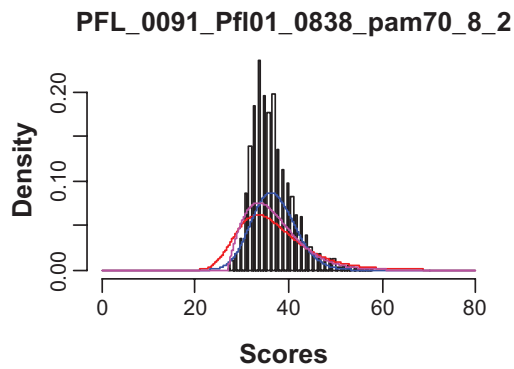
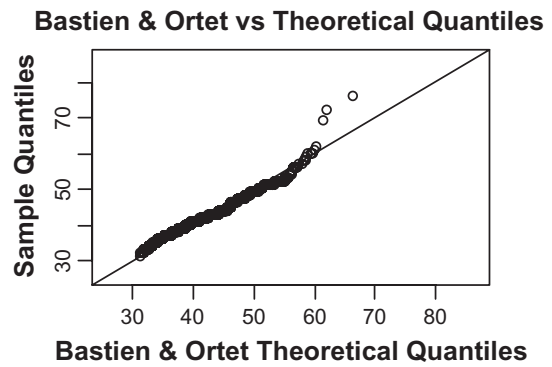
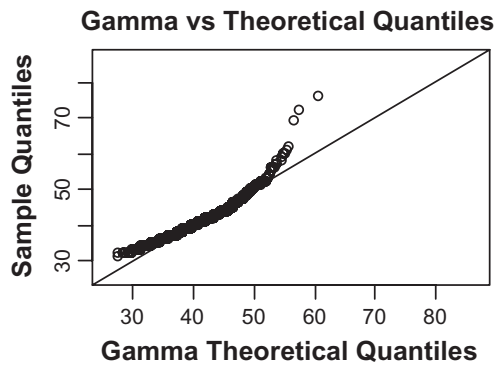
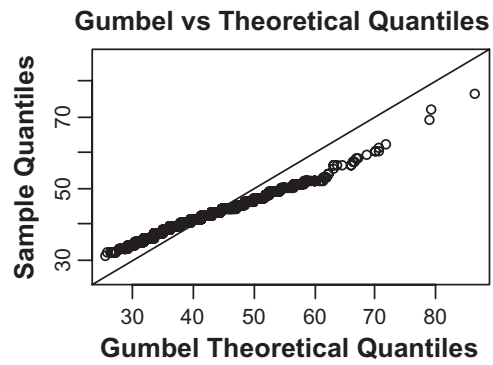
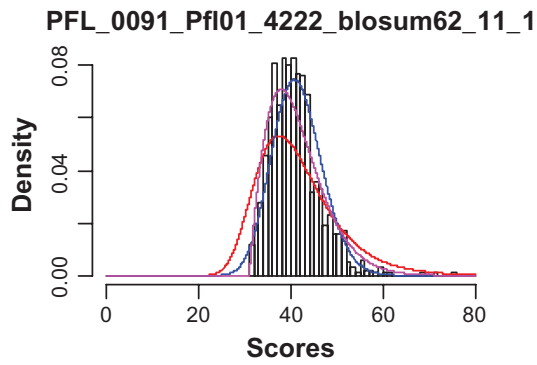


Gamma vs Theoretical Quantiles



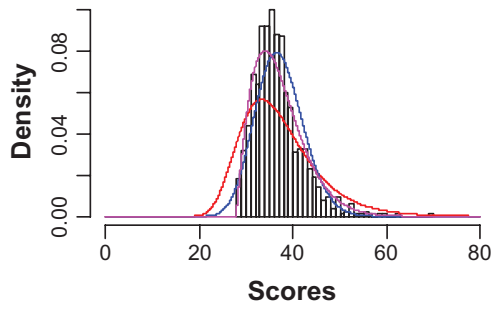
Bastien & Ortet vs Theoretical Quantiles



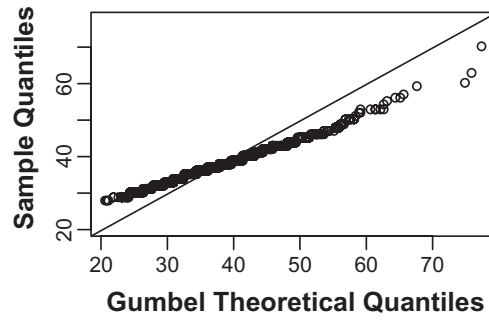




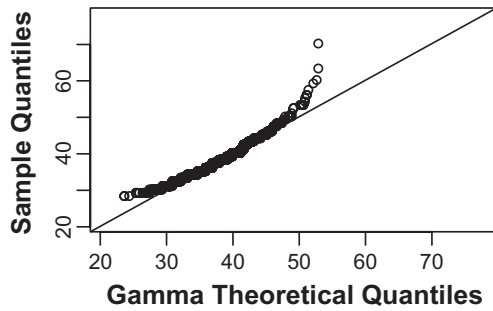
PFL_0091_PfI01_0838_pam70_10_1



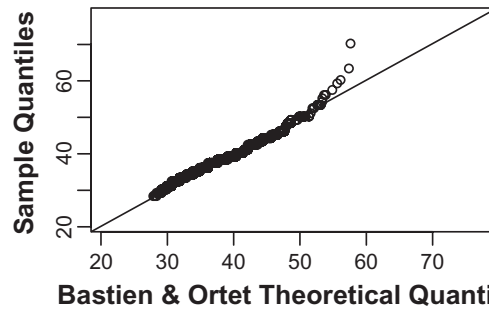
Gumbel vs Theoretical Quantiles



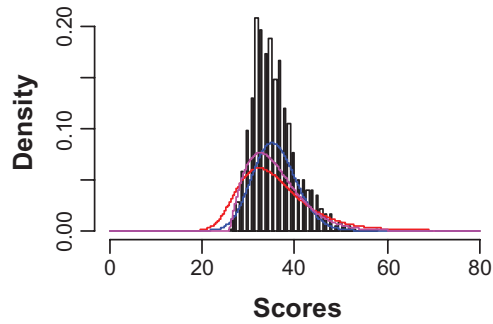
Gamma vs Theoretical Quantiles



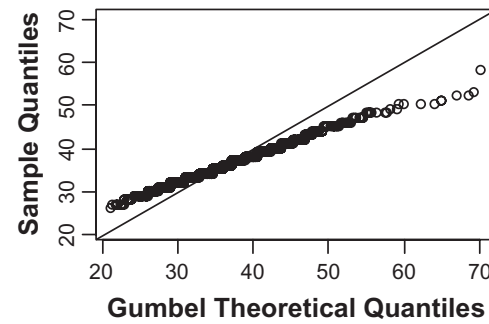
Bastien & Ortet vs Theoretical Quantiles



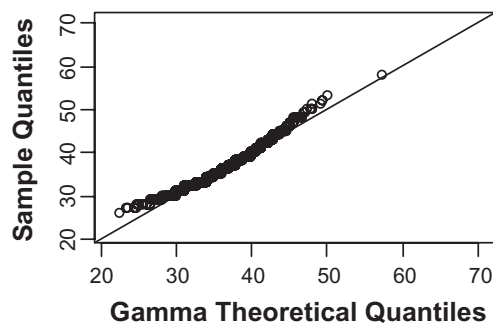
PFL_0091_PfI01_0838_blosum62_9_2



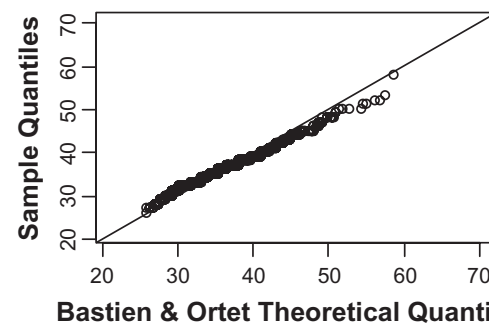
Gumbel vs Theoretical Quantiles



Gamma vs Theoretical Quantiles

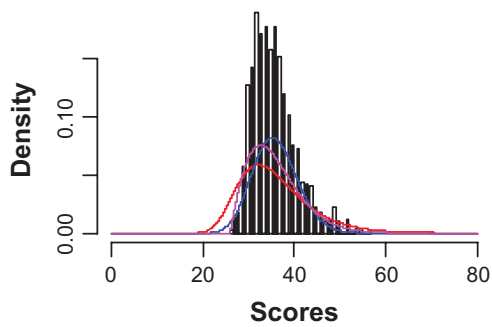


Bastien & Ortet vs Theoretical Quantiles

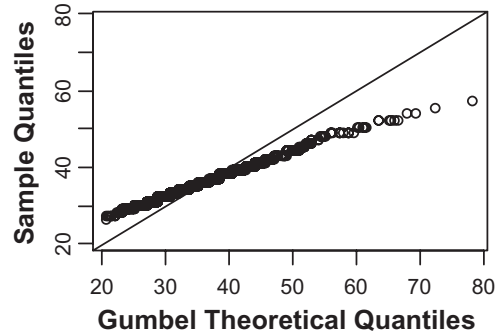




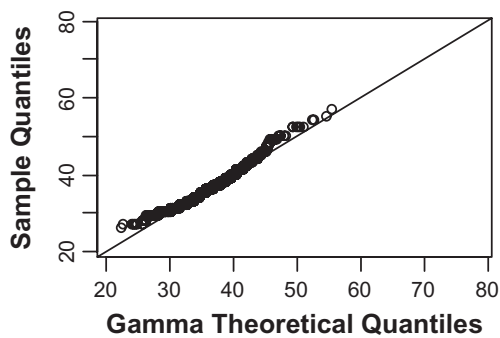
PFL_0091_PfI01_0838_blosum62_11_1



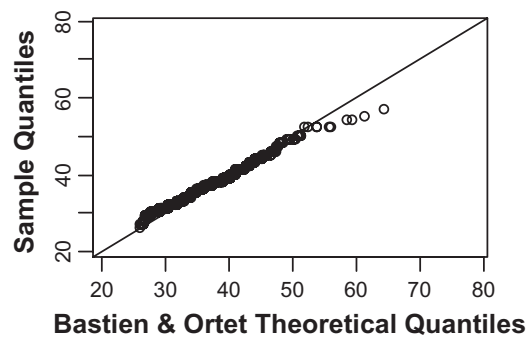
Gumbel vs Theoretical Quantiles



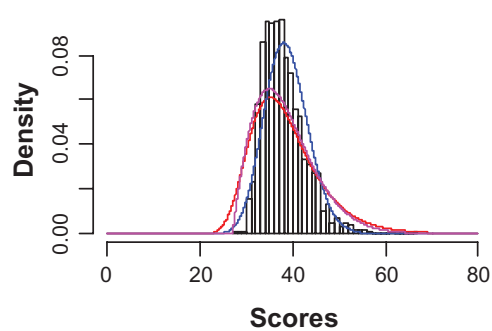
Gamma vs Theoretical Quantiles



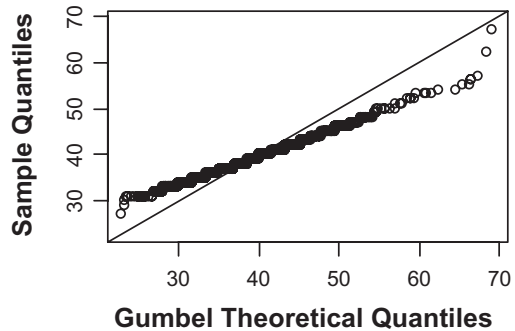
Bastien & Ortet vs Theoretical Quantiles



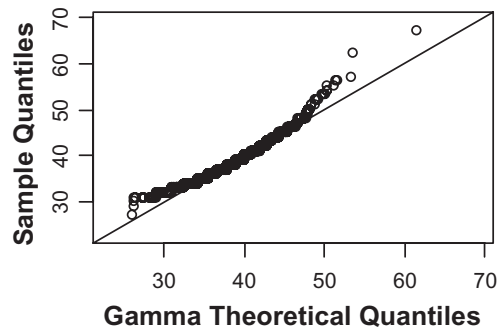
PFL_0091_PfI01_0046_pam70_8_2



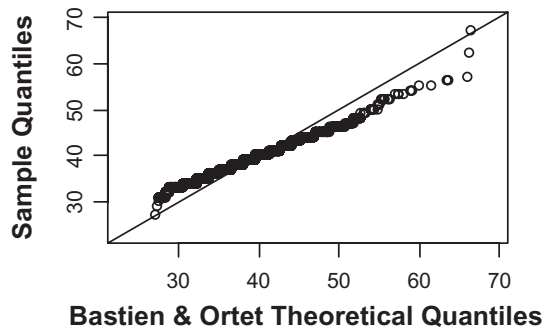
Gumbel vs Theoretical Quantiles



Gamma vs Theoretical Quantiles

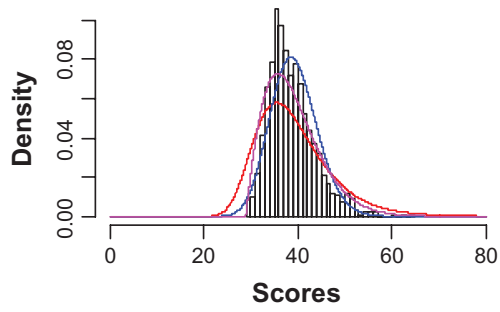


Bastien & Ortet vs Theoretical Quantiles

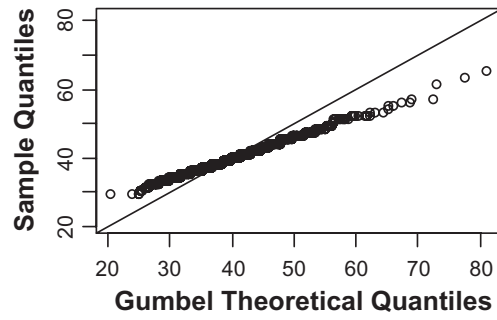




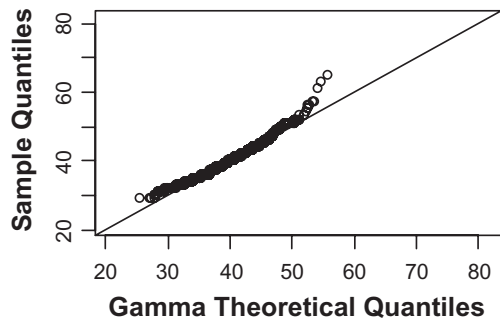
PFL_0091_Pfl01_0046_pam70_10_1



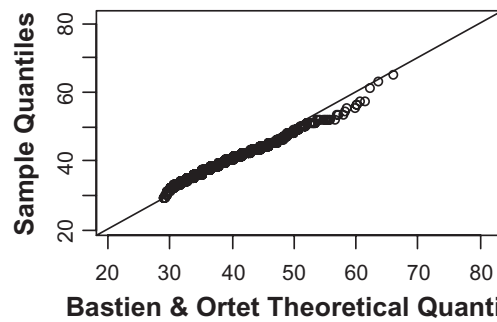
Gumbel vs Theoretical Quantiles



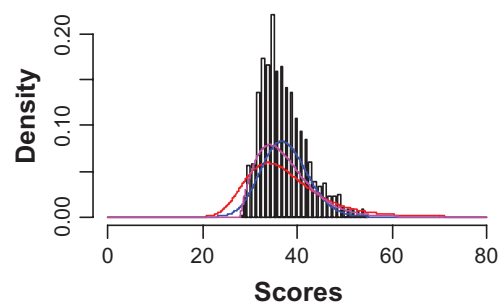
Gamma vs Theoretical Quantiles



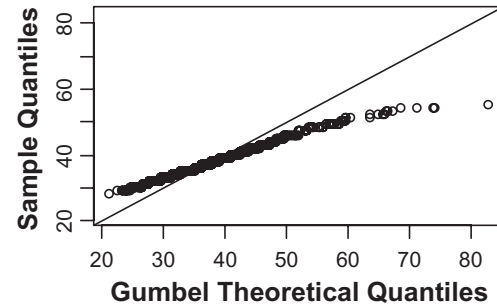
Bastien & Ortet vs Theoretical Quantiles



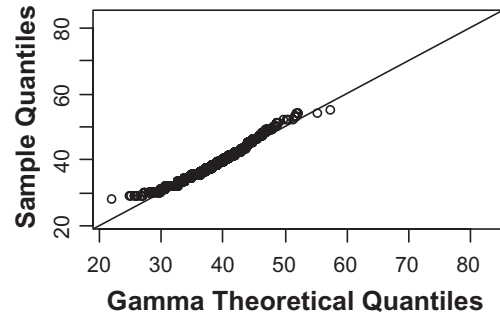
PFL_0091_Pfl01_0046_blosum62_9_2



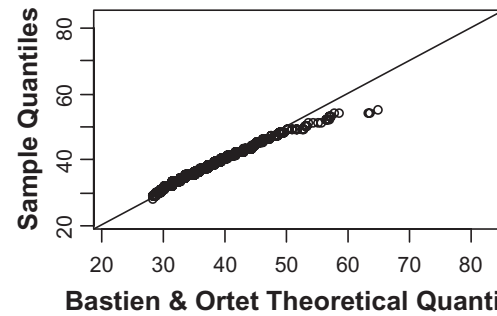
Gumbel vs Theoretical Quantiles



Gamma vs Theoretical Quantiles

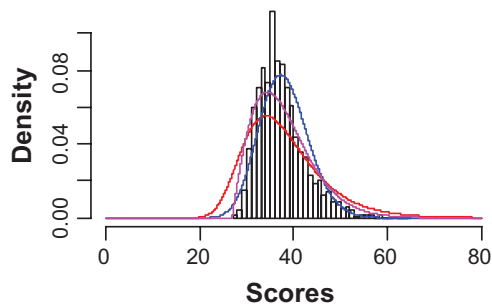


Bastien & Ortet vs Theoretical Quantiles

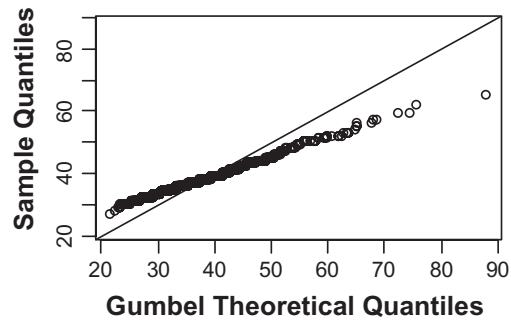




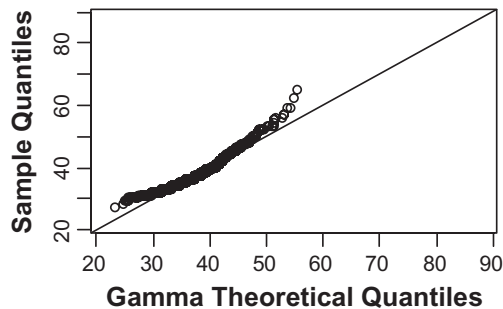
PFL_0091_Pf01_0046_blosum62_11_1



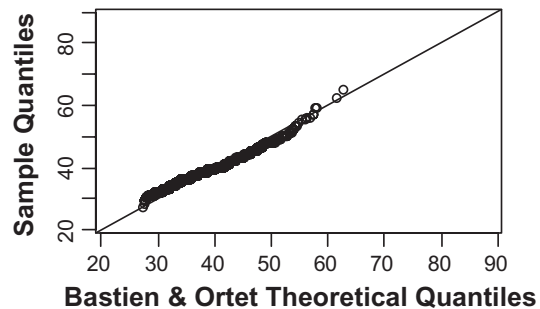
Gumbel vs Theoretical Quantiles



Gamma vs Theoretical Quantiles



Bastien & Ortet vs Theoretical Quantiles



Publish with Libertas Academica and every scientist working in your field can read your article

"I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely."

"The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I've never had such complete communication with a journal."

"LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought."

Your paper will be:

- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

<http://www.la-press.com>