Journal of Computational Biology: http://mc.manuscriptcentral.com/liebert/jcb

Statistical significance of normalized global alignment

Journal:	*Journal of Computational Biology*
Manuscript ID:	JCB-2012-0167.R3
Manuscript Type:	Original Paper
Keyword:	Homology, proteins, SEQUENCE ANALYSIS, algorithms, alignment
Abstract:	The comparison of homologous proteins from different species is a first step towards a function assignment and a reconstruction of the species evolution. Though local alignment is mostly used for this purpose, global alignment is important for constructing multiple alignments or phylogenetic trees. However, statistical significance of global alignments is not completely clear, lacking a specific statistical model to describe alignments or depending on computationally expensive methods like Z-score. Recently we presented a normalized global alignment, defined as the best compromise between global alignment cost and length, and showed that this new technique led to better classification results than Z-score at a much lower computational cost. However, it is necessary to analyze the statistical significance of the normalized global alignment in order to be considered a completely functional algorithm for protein alignment. Experiments with unrelated proteins extracted from the SCOP ASTRAL database showed that normalized global alignment scores can be fitted to a log-Normal distribution. This fact, obtained without any theoretical support, can be used to derive statistical significance of normalized global alignments. Results are summarized in a table with fitted parameters for different scoring schemes. Software used to compute normalized global alignments is available from http://www3.uji.es/~peris/nga

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

jcbnat.bst
NGA_logNormal_coeffs.ods

Statistical significance of normalized global alignment

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Abstract

The comparison of homologous proteins from different species is a first step towards a function assignment and a reconstruction of the species evolution. Though local alignment is mostly used for this purpose, global alignment is important for constructing multiple alignments or phylogenetic trees. However, statistical significance of global alignments is not completely clear, lacking a specific statistical model to describe alignments or depending on computationally expensive methods like Z-score. Recently we presented a normalized global alignment, defined as the best compromise between global alignment cost and length, and showed that this new technique led to better classification results than Z-score at a much lower computational cost. However, it is necessary to

[1] To whom correspondence should be addresed.

analyze the statistical significance of the normalized global alignment in order to be considered a completely functional algorithm for protein alignment.

Experiments with unrelated proteins extracted from the SCOP ASTRAL database showed that normalized global alignment scores can be fitted to a log-Normal distribution. This fact, obtained without any theoretical support, can be used to derive statistical significance of normalized global alignments. Results are summarized in a table with fitted parameters for different scoring schemes.

Software used to compute normalized global alignments is available from http://www3.uji.es/~peris/nga.

Key words: Global alignment, normalization, fractional programming, database search, homologous proteins.

1 Introduction

Sequence alignment lies at the heart of many bioinformatics algorithms, helping with the inference of protein function and structure, homology relationships, and building of evolutionary trees. With the advent of next-generation sequencing machines (see, for example, (baker10)), protein sequence databases are expected to grow faster, but structural and functional properties, experimentally derived from spectrometric and chemical analysis, are more difficult to be confirmed, in spite of recent advances in crystallography techniques ((barty12)). Sequence comparison with previously characterized proteins can help to assign a function to a newly obtained protein.

Pairwise alignments can be divided into two types: global and local. Global alignment (GA) methods compare sequences that are supposed to be similar along their whole length. These methods are used in some multiple alignment algorithms and for building evolutionary trees of sequences with similar lengths. Needleman-Wunsch algorithm ((needleman)) is the most popular GA method, though some heuristic algorithms for long sequences have been presented recently ((bray11)). On the other side, local alignment methods (LA) are used to find similar regions in the strings compared (for example, similar protein domains), and include the rigorous Smith-Waterman algorithm ((smith81)) and other heuristic methods like BLAST ((altschul97)) and FASTA ((fasta88)), more popular for their speed and efficiency. In both cases, an alignment score can be considered as a measure of the shared amount of information (SAI) between a protein and its antecesor ((bastien08)).

Raw scores, obtained with either local or global alignment, depend on sequences length and composition, therefore a high score does not imply a high quality of the alignment and a further study is needed to assess homology of a protein pair. Two major methods for evaluating statistical significance have been proposed. The first method is based on obtaining, either theoretically or experimentally, the scores probability distribution. (karlin90) derived an estimation of the probability of finding a local ungapped alignment between two random sequences a and b with a score $S(a,b)$ greater than s using an Extreme Value Distribution (EVD) or Gumbel distribution type I:

$$P(S(a,b) > s) = 1 - \exp\left(-K \cdot m \cdot n \cdot e^{-\lambda s}\right) \tag{1}$$

where m and n are the sequence lengths (a and b, respectively) and K and λ are constants that depend on the average sequence composition and scoring matrix. This model was theoretically proven for ungapped alignments, though further studies have shown that it is also valid for local gapped alignment statistics ((mott00; altschul01)). Though less attention has been paid to global alignment statistics, (pang05) found that global scores for random protein sequences of similar length could be fitted to a three-parameter Gamma distribution. This distribution was the best fit for real but unrelated sequences of similar length, though in some cases no probability distribution agreed perfectly well to real score distribution.

The second method used for evaluating statistical significance is Z-score ((fasta85)), defined as:

$$Z'(a,b) = \frac{S(a,b) - \mu}{\sigma} \tag{2}$$

Sequence a is shuffled and aligned with sequence b a large number of times (usually, 100–1000 shufflings), and from the resulting scores the mean μ and standard deviation σ are obtained. $Z'(a,b)$ is not a symmetrical score, as it depends on the sequence being shuffled, so (comet99) redefined it as:

$$Z(a,b) = \min\left[Z'(a,b), Z'(b,a)\right] \tag{3}$$

The empirical rule that $Z(a,b) \geq 8$ corresponds to significant alignments was theoretically proven by (bastien04a), though a Z-score probability distribution is needed to assure statistical relevance, particularly for high values in the distribution tail. (comet99) studied Z-score for local alignments and found that quasi-real sequences (shuffled versions of real proteins) followed EVD model, but real proteins showed a deviation in the distribution tail. (webber01) studied global alignment Z-scores obtained from unrelated protein sequences and concluded that the best fit was given by the three-parameter gamma distribution. Parameters obtained for different scoring matrices and length independent gap penalties (gap extension equal to zero) were provided, so that P-values can be obtained for these scoring schemes. These results are based on experimentally fitted distributions alignment scores, using either real proteins, or quasi-real or random sequences, but with no theoretical justification. However, some studies ((bastien08; bastien08b)) derive the EVD distribution for local and global alignment score, based on the assumption of basic processes guiding biological evolution of proteins.

Recently we proposed a normalized global alignment (NGA) score that corrected the length dependence of a raw global score ((peris11)). NGA score was shown to be linearly related to Z-score, and computational cost was sensibly lower, allowing for a faster detection of homologous proteins. In this work, NGA statistical significance is studied, and applied to obtain P-values of normalized global alignments of protein pairs.

2 Approach

For an arbitrary global alignment E between two biological sequences a and b (of lengths m and n, respectively), we can compute its score $S(E)$ using a substitution matrix σ (that assigns a value to every residue pair) and an affine function that penalizes gaps (with g_o for opening a new gap, and g_e for each additional gap extension). Considering the set of all possible alignments $\mathcal{E}_{ab}$, the normalized global alignment is defined as ((peris11))

$$NGA(a,b) = \max_{E \in \mathcal{E}_{ab}} \frac{S(E)}{L(E)}, \tag{4}$$

where $L(E)$ is the alignment length.

This optimization problem cannot be solved using a traditional dynamic programming technique. Solution of problem (4) involves optimizing the ratio of two linear functions, and this question was solved by Dinkelbach ((dinkel67)) applying the so-called fractional programming technique. This algorithm has been previously used to compute normalized edit distances ((vidal95; marzal93; peris09)). Following Dinkelbach's solution, the alignment that maximizes the ratio of alignment score and length can be obtained by solving the parametrized equation:

$$\hat{d}(\rho) = \max_{E \in \mathcal{E}_{ab}} (S(E) - \rho L(E)) \qquad L(E) > 0, \forall E \in \mathcal{E}_{ab}. \tag{5}$$

where ρ is the parameter to be found. This equation is iteratively solved by starting with an initial guess for ρ and obtaining the alignment that maximizes $\hat{d}(\rho)$. This alignment is used to compute a new approximation for ρ, and the procedure is repeated until ρ converges to NGA score,

$$\rho_{i+1} = \frac{S(E_i)}{L(E_i)}, \tag{6}$$

where E_i represents the best alignment optimized in Equation (5) using ρ_i.

Though different starting values for ρ_0 can be used, we take the standard global alignment E_{GA} score divided by the alignment length, the so-called post-normalized global alignment (PNGA),

$$\rho_0 = \frac{S(E_{GA})}{L(E_{GA})} \tag{7}$$

Maximization problem (5) can be posed in terms of the classical global alignment problem:

$$\begin{aligned} \max_{E \in \mathcal{E}_{ab}} (S(E) - \rho L(E)) &= \max_{E \in \mathcal{E}_{ab}} (S(E) - \rho(m + n - n_s)) \\ &= \max_{E \in \mathcal{E}_{ab}} (S(E) + \rho n_s) - \rho(m + n)) \end{aligned} \tag{8}$$

where n_s is the number of aligned letter-pairs in E.

Solution to problem (5) can now be related to global alignment, just adding parameter ρ to every value in the scoring matrix in every iteration or, equivalently, modifying the diagonal terms in the programming dynamic matrix adding ρ, so that a parametrized global alignment $GA_\rho(a, b)$ is obtained,

$$\max_{E \in \mathcal{E}_{ab}} (S(E) - \rho L(E)) = GA_\rho(a, b) - \rho(m + n) \tag{9}$$

Computational cost of NGA algorithm is $\mathcal{O}(smn)$, depending on the number of iterations s. In previous studies with either normalized edit distance ((peris09)) or NGA ((peris11)), convergence was reached with an average of 2.5 iterations, never exceeding a value of 5. Anyway, Dinkelbach's algorithm is guaranteed to converge given a finite set of alignments $\mathcal{E}_{ab}$ ((dinkel67)).

Properties of NGA scores were extensively studied in ((peris11)) for a set of proteins extracted from ASTRAL database ((astral)). It was found a linear relationship between NGA score and Z-score (see Fig. 1), so that all protein pairs with a NGA score over some threshold were homologous and with a high Z-score value. This linearity held even for proteins with a biased amino acid composition, such as those of *Plasmodium falciparum*. Furthermore, NGA improves Z-score performance on homologous detection in a database (Fig. 2). It was shown that NGA leads to suboptimal alignments with a better overlap between the proteins compared.

[Figure 1 about here.]

[Figure 2 about here.]

However, obtaining a high NGA score does not ascertain that the two proteins are homologous. In order to decide if a high score may be obtained by chance or it points to a relationship between proteins, the score probability distribution of unrelated protein pairs must be studied.

3 Methods

3.1 Software

NGA algorithm was implemented in C++ in (peris11). Basically, it is a modification of Needleman-Wunsch algorithm ((needleman)), changing diagonal contributions in edit graph to include ρ value in Eq. (9). The alignments were not penalized on ending gaps. Z-score was also implemented in the same package, with a standard number of 100 shufflings for each sequence.

3.2 Databases

In this work, a protein subset from ASTRAL database version 1.75 ((astral)) was used. This database is assembled from SCOP database ((hubbard97)), that contains protein domains manually classified according to its hierarchical evolutionary relationships into classes, folds, superfamily and family. Protein domains with similar structures, functions and sequences are arranged in the same family. Members of different families but a common evolutionary relationship are classified into the same superfamily, so they are considered homologous. A fold is comprised of different superfamilies that share common secondary structures, so domains included in different folds are supposed to be non homologous. Homology of sequence pairs in the same fold and different superfamily is not clear, so they are not considered in our experiments.

ASTRAL SCOP database is used as a benchmark to evaluate the ability of algorithms to detect remote homologous, because the protein domains are filtered so that every pair does not exceed a predefined sequence identity percentage. In this work 40% filtered set (astral40) was used and it was further divided into training and test sets.

To obtain domain sets of similar sizes, the training set contains the odd numbered folds in classes a, c, e and g, and the even numbered folds in classes b, d and f, while the rest of the protein domains were included in the test set ((price05)). The test set was used in (peris11) for classification experiments. The training set was further filtered selecting randomly one sequence from each fold (592 sequences), so that all protein pairs are non homologous.

3.3 Scoring schemes

For experiments we used classical scoring matrices from BLOSUM series (BLOSUM50 and BLOSUM62 – (blosum)) and PAM series (PAM120 and PAM250 – (dayhoff78)). BLOSUM matrices are widely used in database search, while PAM matrices were chosen because they represent different evolutionary time. Though only length independent gap penalties were used in the previous paper ((peris11)), general affine gap penalty functions were used in this study, with several matrix/gap penalty combinations (see Table 1).

3.4 Statistics

All the curve fitting to different distributions was computed with the R statistical package ((R)).

4 Results

NGA scores and Z-scores obtained for every scoring scheme were fitted to several positive skewed density distributions (Weibull, log-Normal, Extreme Value Distribution and three-parameter Gamma), and the Kolmogorov-Smirnov test was performed for each fitted distribution.

Our results for Z-score agree with (webber01), who pointed out that Z-score fitted well to several distributions, including log-Normal and three-parameter Gamma distribution, though the last one was chosen based only on statistical fitting. As stated by Firth ((firth88)), analyzing log-Normal data assuming a gamma distribution is more efficient than analyzing gamma data assuming log-normality.

The three-parameter Gamma distribution, or Pearson type III distribution, is given by the equation:

$$f(x) = \frac{(x-\xi)^{\alpha-1}}{\Gamma(\alpha)\lambda^{\alpha}} e^{-(x-\xi)/\lambda} \tag{10}$$

where ξ is the location parameter ($0 \leq (x-\xi) < \infty$), α is the shape parameter ($\alpha > 0$) and λ is the scale parameter ($\lambda > 0$).

Log-Normal distribution is given by the equation:

$$f(x) = \frac{\alpha}{\sqrt{2\pi}(x-\xi)} exp\left[-\frac{\left(\ln\left(\frac{x-\xi}{\lambda}\right)^{\alpha}\right)^{2}}{2}\right] \tag{11}$$

where parameters have the same meaning as in Eq. (10). It is convenient to write Eq. (11) as:

$$f(x) = \frac{1}{\sqrt{2\pi}(x-\xi)\sigma} exp\left[-\frac{1}{2}\left(\frac{\ln(x-\xi)-\mu}{\sigma}\right)^{2}\right] \tag{12}$$

where μ and σ are the mean and standard deviation of $\ln(x - \xi)$.

For NGA scores we find similar trends, that is, log-Normal and Gamma are the best fitting distributions according to Kolmogorov-Smirnov and χ-squared tests. However, in this case log-Normal outperforms Gamma distribution, with a lower Kolmogorov-Smirnov mean value. Furthermore, distribution tail fits better to log-Normal equation, as shown in Fig. 3, and Q-Q plots confirm this fact (Fig. 4 and 5). So, we use log-Normal distribution to model NGA data, though this is based only in goodness of fit and with no theoretical support. Furthermore, it must be stated that no scores have been obtained for the high distribution tail (with P-values around 1×10^{-8}), which is a real interesting area for database searching. Though some techniques are available to obtain such a high scores using randomly generated proteins ((bundschuh02; sheetlin05)), there are no such algorithms for real proteins. This problem become worse due to the fact that a normalized score is upper bounded.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

In Table 1 the probability density function (PDF) parameters of log-Normal distribution obtained for every scoring scheme are shown. Fig. 6–9 show the fitted distributions for the different substitution matrices used, and different gap penalty schemes. In the insets the upper tails of the distributions are highlighted because in this area is more difficult to distinguish between homologous and non homologous protein pairs. It can be seen there is a good agreement between experimental and fitted distribution in this twilight region.

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

NGA scores needed to obtain a predetermined P-values of 1×10^{-5} and $1 \times 10^{-8}$are shown in the last columns of Table 1.

[Table 1 about here.]

5 Discussion

In a previous paper we introduced the definition of normalized global alignment, based on the normalized edit distance, as a way to correct the length dependence of a global alignment score. It was shown that NGA allows the selection of a suboptimal global alignment with a better overlap between the proteins compared, which is consistent with the fact that structurally accurate alignments are often suboptimal ((zuker91)). Comparing NGA scores with Z-scores, it was found a linear relationship between both scores. This relationship allows to detect homologous with a high NGA value at a lower computational cost than Z-score. Furthermore, NGA score improves homologous detection on a database search *versus* global aligment score and Z-score.

In this paper we have studied statistical significance of NGA scores, finding that these scores can be fitted experimentally to a three-parameter log-Normal distribution. This distribution allows a good fitting in the right tail of the distribution, an important area to decide if a protein pair is homologous or not. We have obtained the PDF parameters for several scoring schemes. Considering all, we consider that NGA is a good and cheap algorithm to detect homologous protein pairs, or as a first filter to more elaborate and expensive algorithms. Anyway, we must remember that distributions have been obtained considering only goodness of fit and with no theoretical support. Furthermore, it was no possible to obtain high scores with P-values around 1×10^{-8}.

6 Acknowledgements

The authors would like to thank the Editor-in-Chief and the Associate Editors for carefully reading the paper and for their comments which greatly improved the paper.

7 Author Disclosure Statement

No competing financial interests exists.

References

[R] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[altschul01] S.F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa. The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res., 26:351–361., 2001.

[altschul97] S.F. Altschul, T. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res., 25:3389–3402., 1997.

[astral] Steven E. Brenner, Patrice Koehl, and Michael Levitt. The astral compendium for protein structure and sequence analysis. Nucleic Acids Research, 28(1):254–256, 2000.

[baker10] M. Baker. Next-generation sequencing: adjusting to data overload. Nature Methods, 7(7):495–499, 2010.

[barty12] A. Barty, C. Caleman, A. Aquila, N. Timmenu, T.A. Lomb, L. White, J. Andreasson, D. Arnlund, S. Bajt, T.R.M. Barends, M. Barthelmess, M.J. Bogan, C. Bostedt, J.D. Bozek, R. Coffee, N. Coppola, J. Davidsson, D.P. DePonte, R.B. Doak, T. Ekeberg, V. Elser, S.W. Epp, B. Erk, H. Fleckenstein, L. Foucar, P. Fromme, H. Graafsma, L. Gumprecht, J. Hajdu, C.Y. Hampton, R. Hartmann, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M.S. Hunter, L. Johansson, S. Kassemeyer, N. Kimmel,

R.A. Kirian, F Liang, M.and Maia, E. Malmerberg, S. Marchesini, A.V. Martin, K. Nass, R. Neutze, C. Reich, D. Rolles, B. Rudek, A. Rudenko, H. Scott, I. Schlichting, J. Schulz, M.M. Seibert, R.L. Shoeman, R.G. Sierra, H. Soltau, J. Spence, F. Stellato, S. Stern, L. Struder, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, C.B. Wunderer, and H.N. Chapman. Self-terminating diffraction gates femtosecond x-ray nanocrystallography measurements. Nature Photonics, 6:35–40, 2012.

[bastien04a] O. Bastien, J.-C. Aude, S Roy, and E Maréchal. Fundamentals of massive automatic pairwise alignments of protein sequences: Theoretical significance of z-value statistics. Bioinformatics, 20:534–537., 2004.

[bastien08] O. Bastien and E. Maréchal. Evolution of biological sequences implies an extrema value distribution of type i for both global and local pairwise alignments scores. BMC Bioinformatics, 9:332, 2008.

[bastien08b] O. Bastien. A simple derivation of the distribution of pairwise local protein sequence alignment scores. Evolutionary Bioinformatics, 4:41–45, 2008.

[blosum] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Science USA, 89:10915–10919., 1992.

[bray11] N. Bray, I. Dubchak, and Pachter L. Avid: A global alignment program. Genome Research, 13:97–102, 2011.

[bundschuh02] R. Bundshuh. Rapid significance estimation in local alignment with gaps. Journal of Computational Biology, 9(2):243–260., 2002.

[comet99] J.P. Comet, J.C. Aude, E. Glémet, A. Wozniak, J.L. Risler, A. Hénaut, and P.P. Slonimski. Significance of z-value statistics of smith-waterman scores for protein alignments. Computers and Chemistry, 23:317–331., 1999.

[dayhoff78] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. Matrices for detecting distant relationships, volume 5, pages 345–358. National Biomedical Research Fundation, Washington DC, Washington, 1978.

[dinkel67] W. Dinkelbach. On nonlinear fractional programming. Management Science, 18(7):492–498., 1967.

[fasta85] D. Lipman and W. Pearson. Rapid and sensitive protein similarity searches. Science, 227:1435–1441., 1985.

[fasta88] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. Proceedings of the National Academy of Science USA, 85:2444–2448., 1988.

[firth88] D. Firth. Multiplicative errors: Log-normal or gamma? J. R. Statist. Soc. B, 50(2):266–268, 1988.

[hubbard97] Tim J. P. Hubbard, Alexey G. Murzin, Steven E. Brenner, and Cyrus Chothia. Scop: a structural classification of proteins database. J. Mol. Biol, 247:536–540, 1997.

[karlin90] S Karlin and S.F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA, 87:2264–2268., 1990.

[marzal93] A. Marzal and E. Vidal. Computation of normalized edit distances and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(9):926–932., 1993.

[mott00] R. Mott. Accurate formula for p-value of gapped local sequence and profile alignment. Journal of Molecular Biology, 300:649–659., 2000.

[needleman] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48:443–453., 1970.

[pang05] H. Pang, J. Tang, S. Chen, and S. Tao. Statistical distributions of optimal global alignment scores of random protein sequences. BMC Bioinformatics, 6:257, 2005.

[peris09] Guillermo Peris and Andrés Marzal. A screening method for z-value assessment based on the normalized edit distance. In IWANN '09: Proceedings of the 10th International Work-Conference on Artificial Neural Networks, pages 1154–1161, 2009.

[peris11] G. Peris and A. Marzal. Normalized global aligment for protein sequences. Journal of Theoretical Biology, 291C:22–28, 2011.

[price05] Gavin A. Price, Gavin E. Crooks, Richard E. Green, and Steven E. Brenner. Statistical evaluation of pairwise protein sequence comparison with the bayesian bootstrap. Bioinformatics, 21(20):3824–3831, 2005.

[sheetlin05] S. Sheetlin, Y. Park, and JL. Spouge. The gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. Nucleic Acids Res., 33(15):4987–4994., 2005.

[smith81] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, 147(1):195–197, 1981.

[vidal95] E. Vidal, A. Marzal, and P. Aibar. Fast computation of normalized edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(9):899–902., 1995.

[webber01] Caleb Webber and Geoffrey J. Barton. Estimation of p-values for global alignments of protein sequences. Bioinformatics, 17(12):1158–1167, 2001.

[zuker91] M Zuker. Suboptimal sequence alignment in molecular biology: alignment with error analysis. Journal of Molecular Biology, 221(2):403–420, 1991.

List of Figures

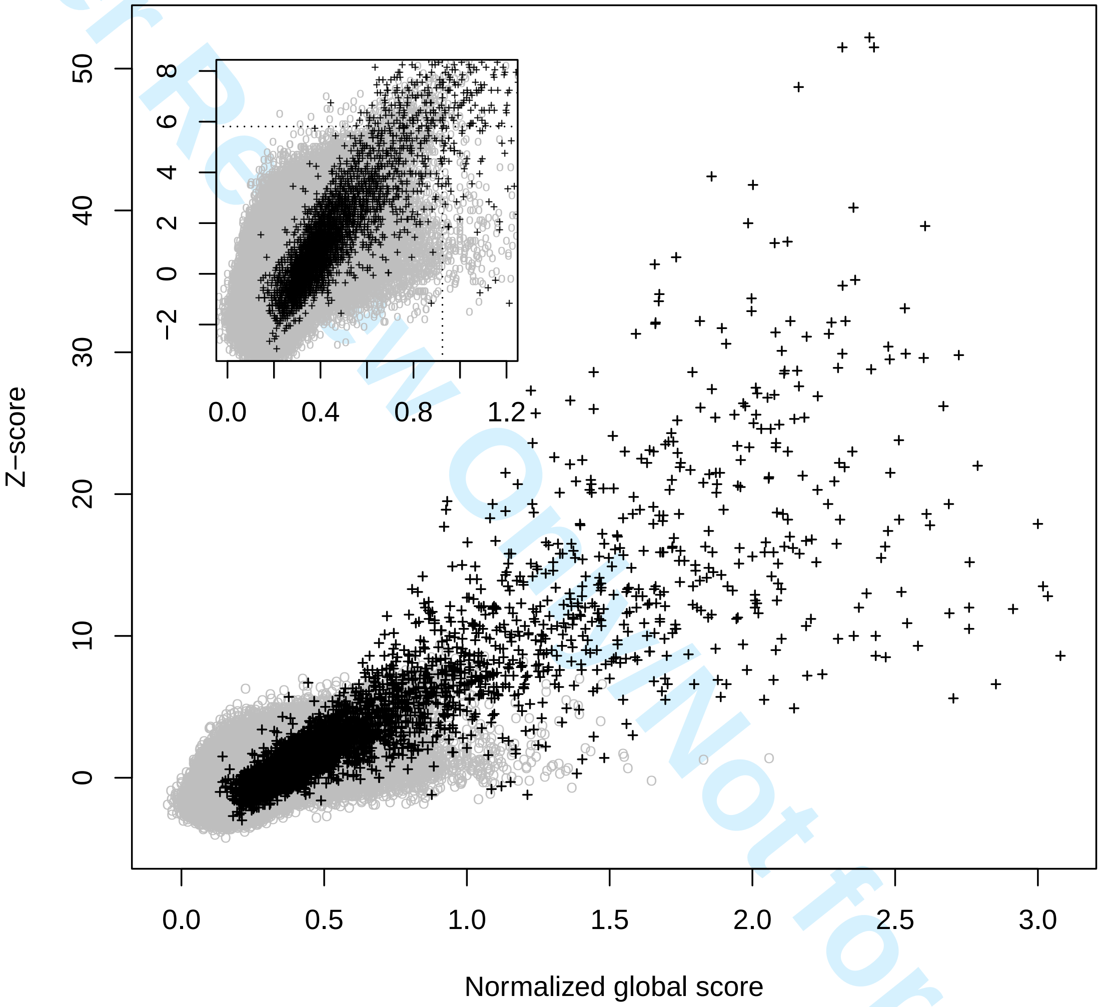

Figure 1: Classification experiments with scoring matrix BLOSUM50 and gap penalties $g_o = 12$ and $g_e = 0$, using astral40 training data set. *Z*-score *vs.* NGA plot is shown, where black crosses represent homologous protein pairs, and grey circles non homologous pairs.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 1 (of 9)

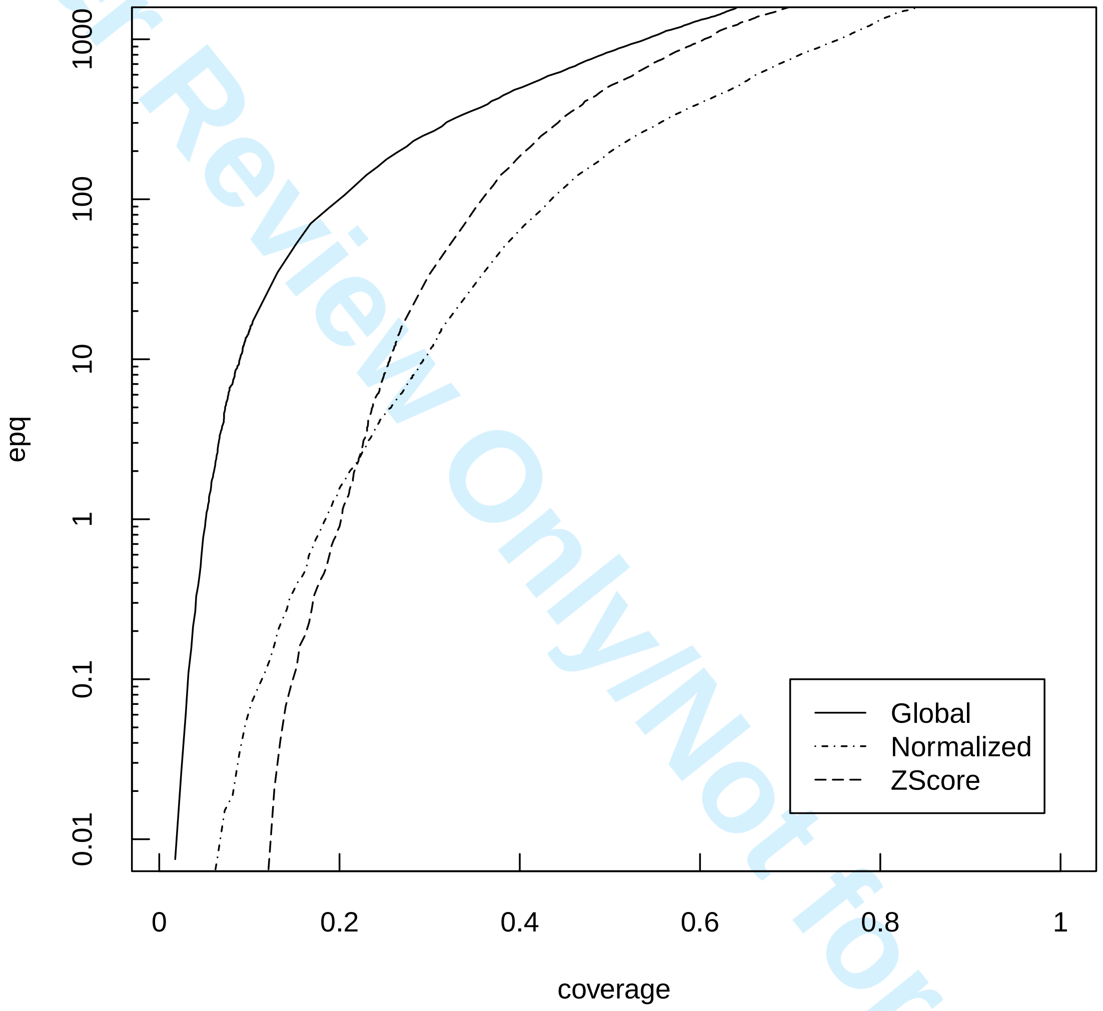

Figure 2: Classification experiments with scoring matrix BLOSUM50 and gap penalties $g_o = 12$ and $g_e = 0$, using astral40 training data set. Error per query *vs.* coverage plot is shown using global alignment, Z-score and normalized global alignment.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 2 (of 9)

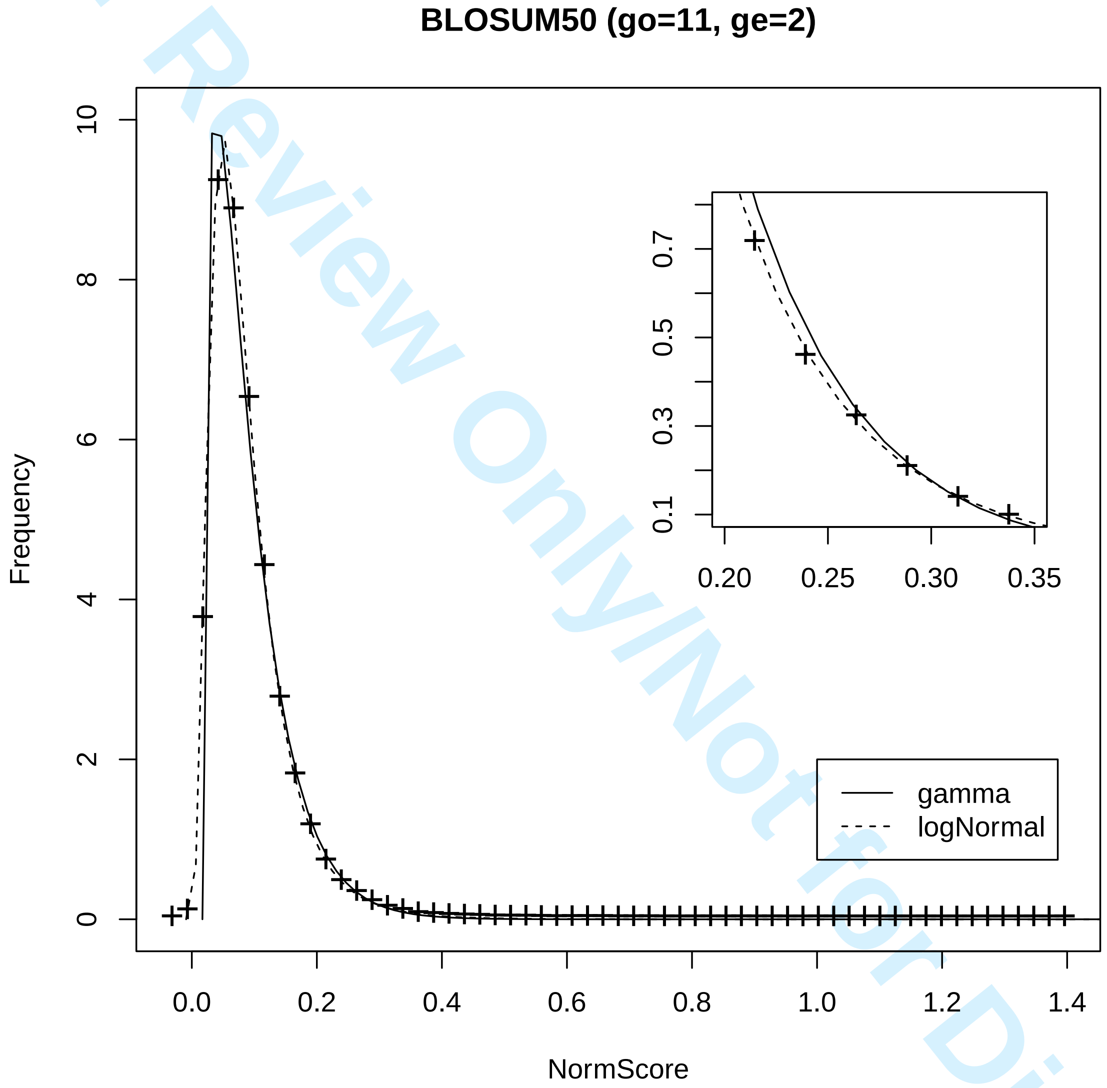

Figure 3: Distribution of NGA scores for astral40 training set, using as a scoring scheme BLOSUM50 and 11/2 for gap penalty. Fitted log-Normal distribution (dotted line) and gamma distribution (solid line) are shown.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 3 (of 9)

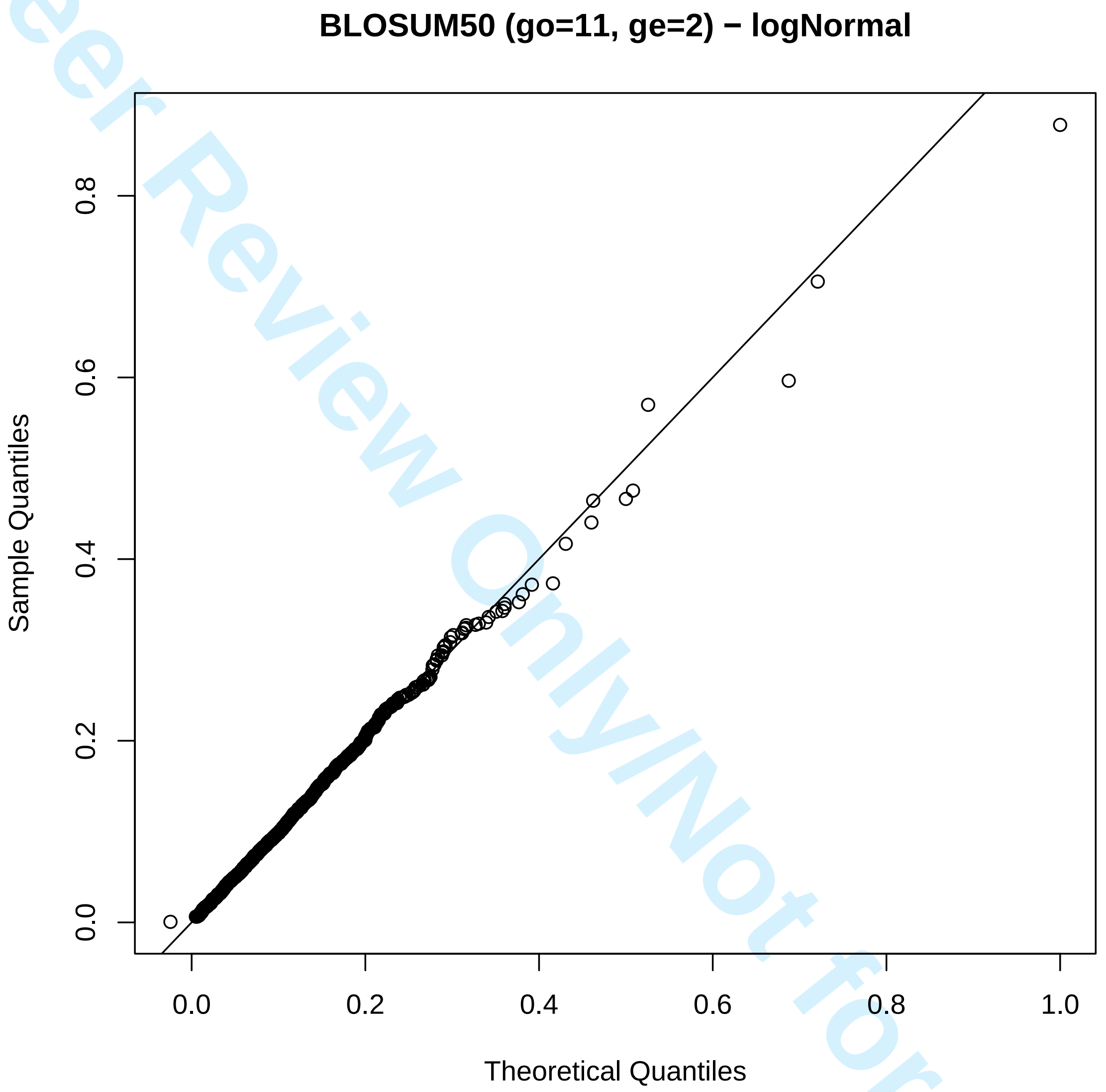

Figure 4: Q-Q plot for log-Normal distribution using the scoring scheme BLOSUM50, 11/2.

$\Uparrow$

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 4 (of 9)

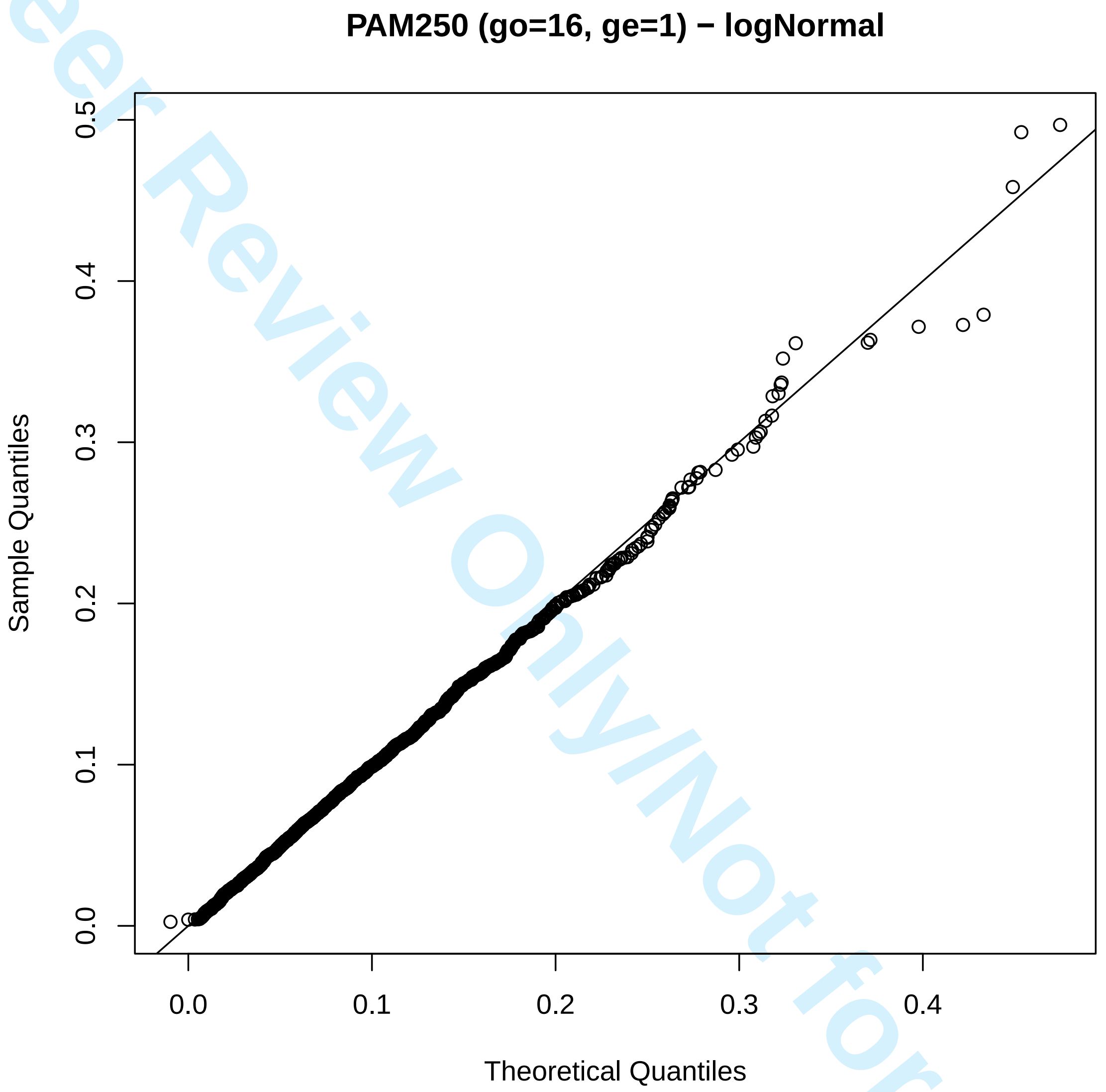

Figure 5: Q-Q plot for log-Normal distribution using the scoring scheme PAM250, 16/1.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 5 (of 9)

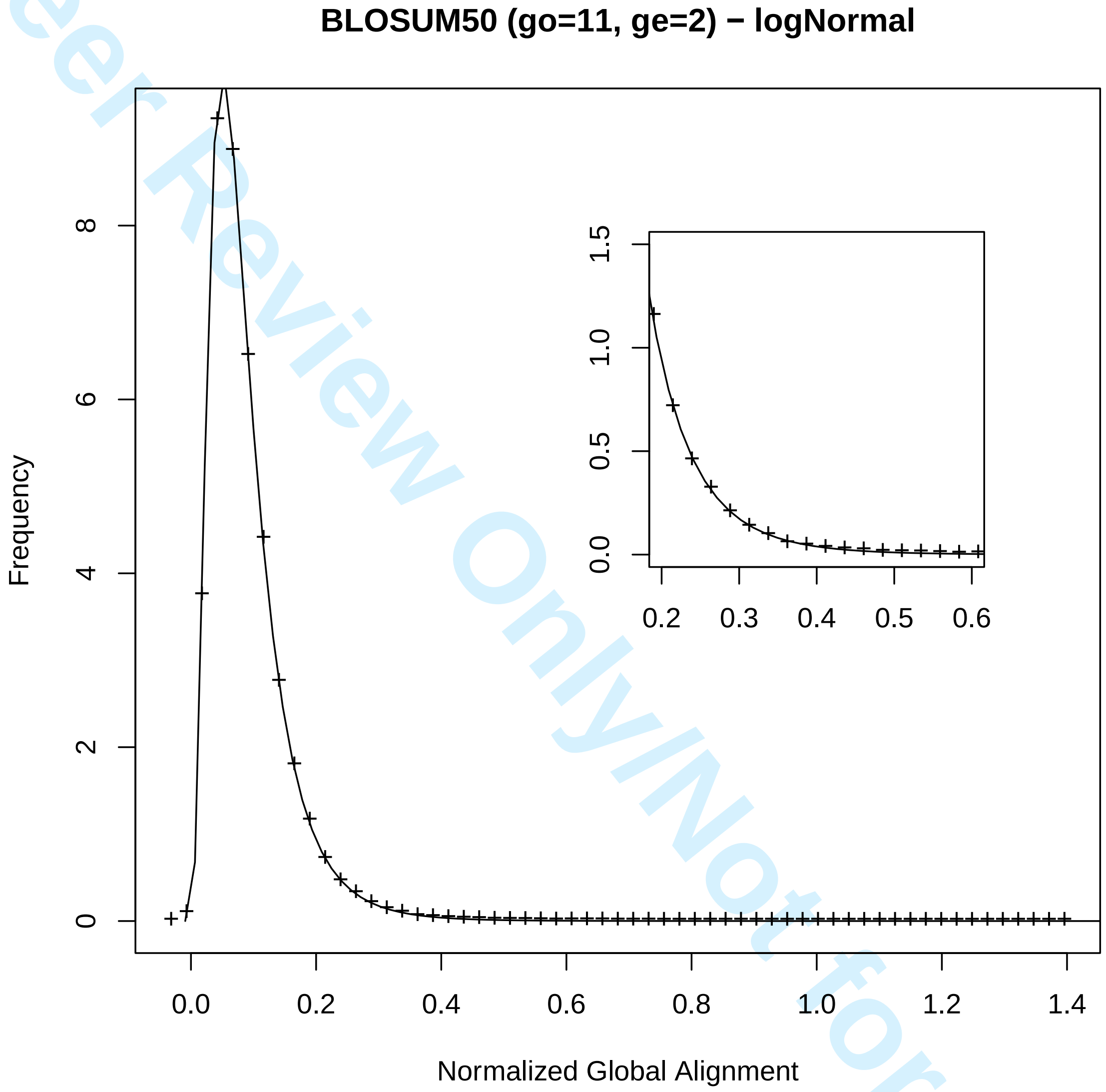

Figure 6: Distribution of NGA scores for astral40 training set, fitted to log-Normal distribution and scoring scheme BLOSUM50, 11/2.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 6 (of 9)

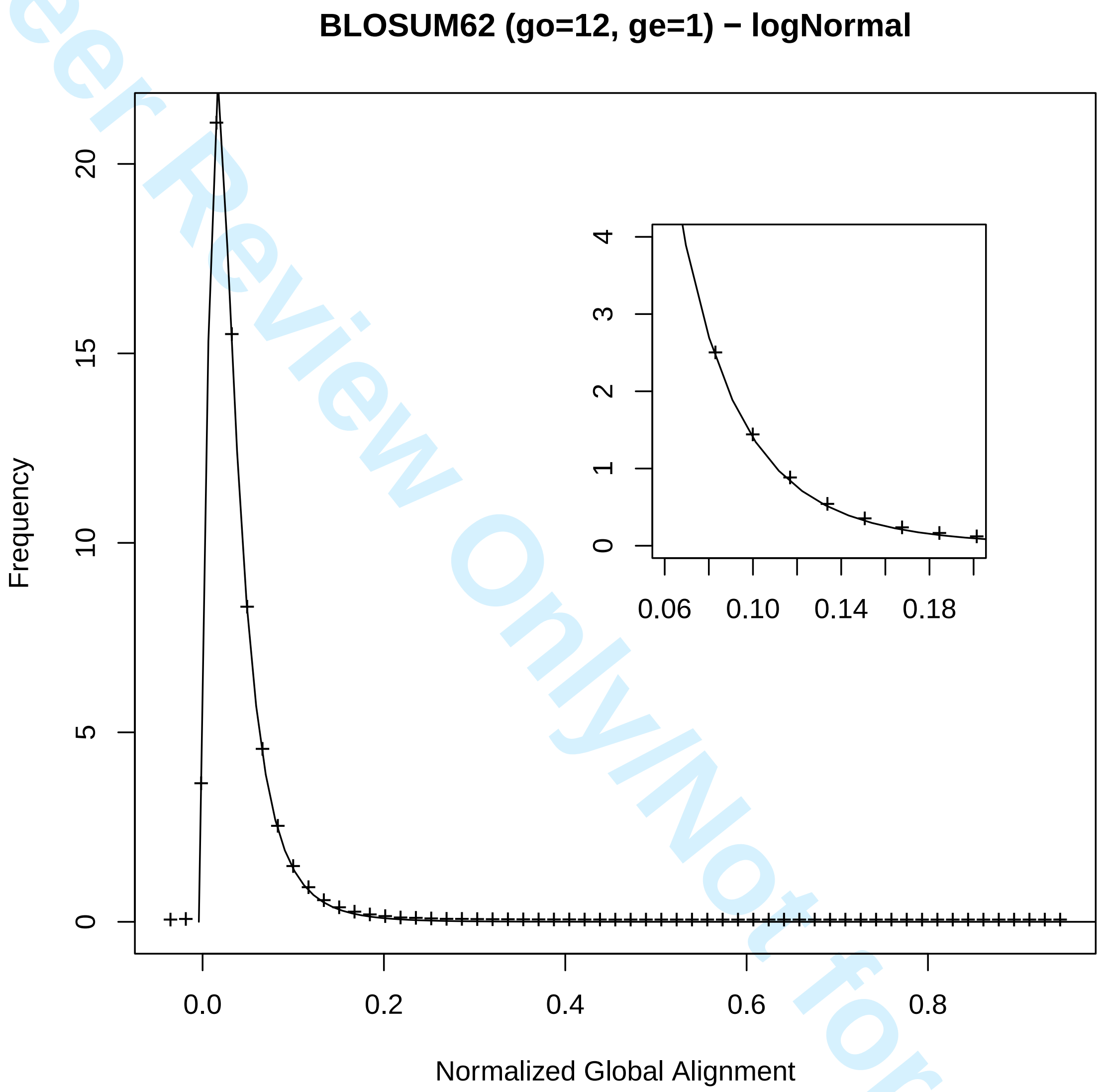

Figure 7: Distribution of NGA scores for astral40 training set, fitted to log-Normal distribution and scoring scheme BLOSUM62, 12/1.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 7 (of 9)

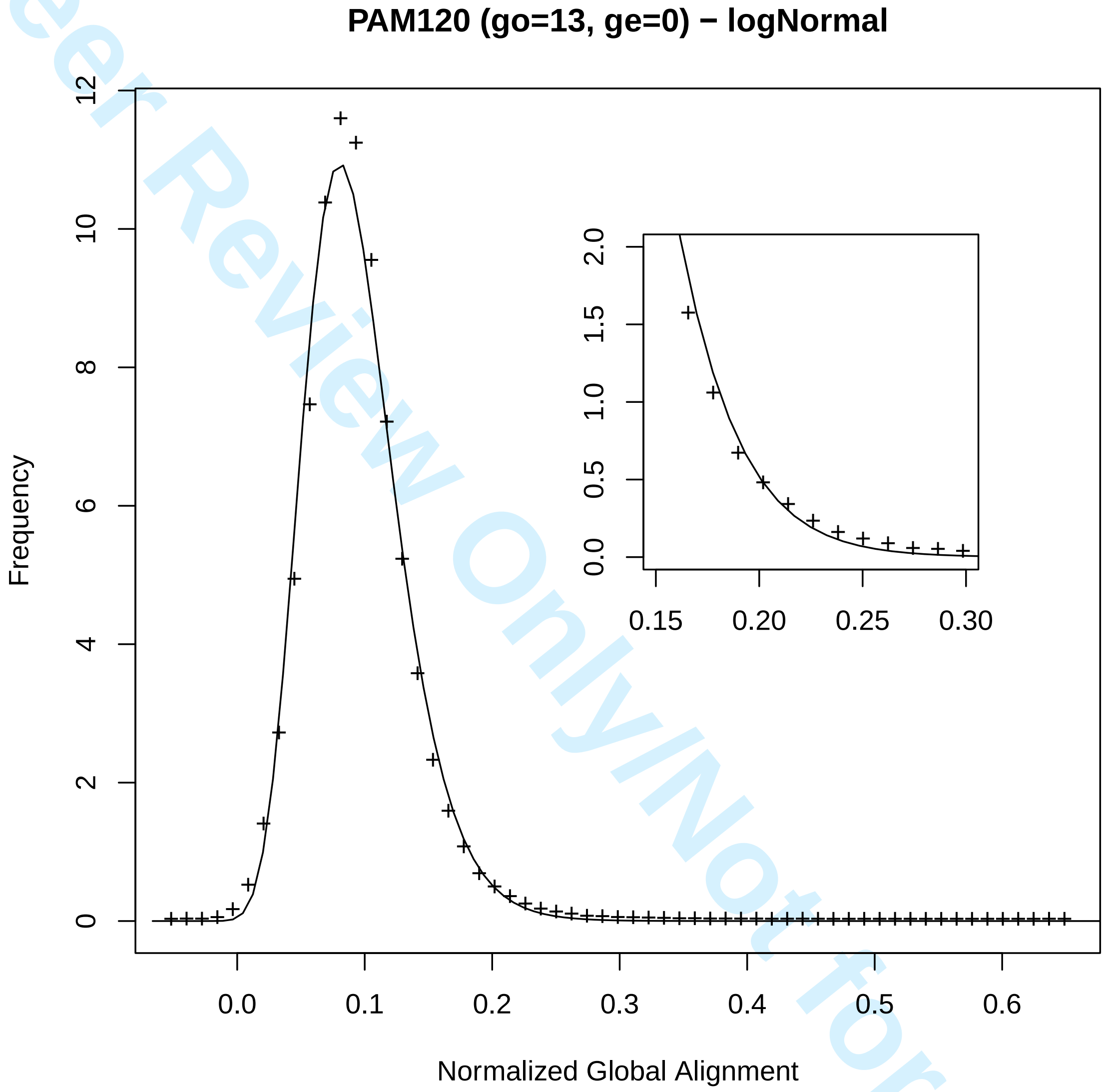

Figure 8: Distribution of NGA scores for astral40 training set, fitted to log-Normal distribution and scoring scheme PAM120, 13/0.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 8 (of 9)

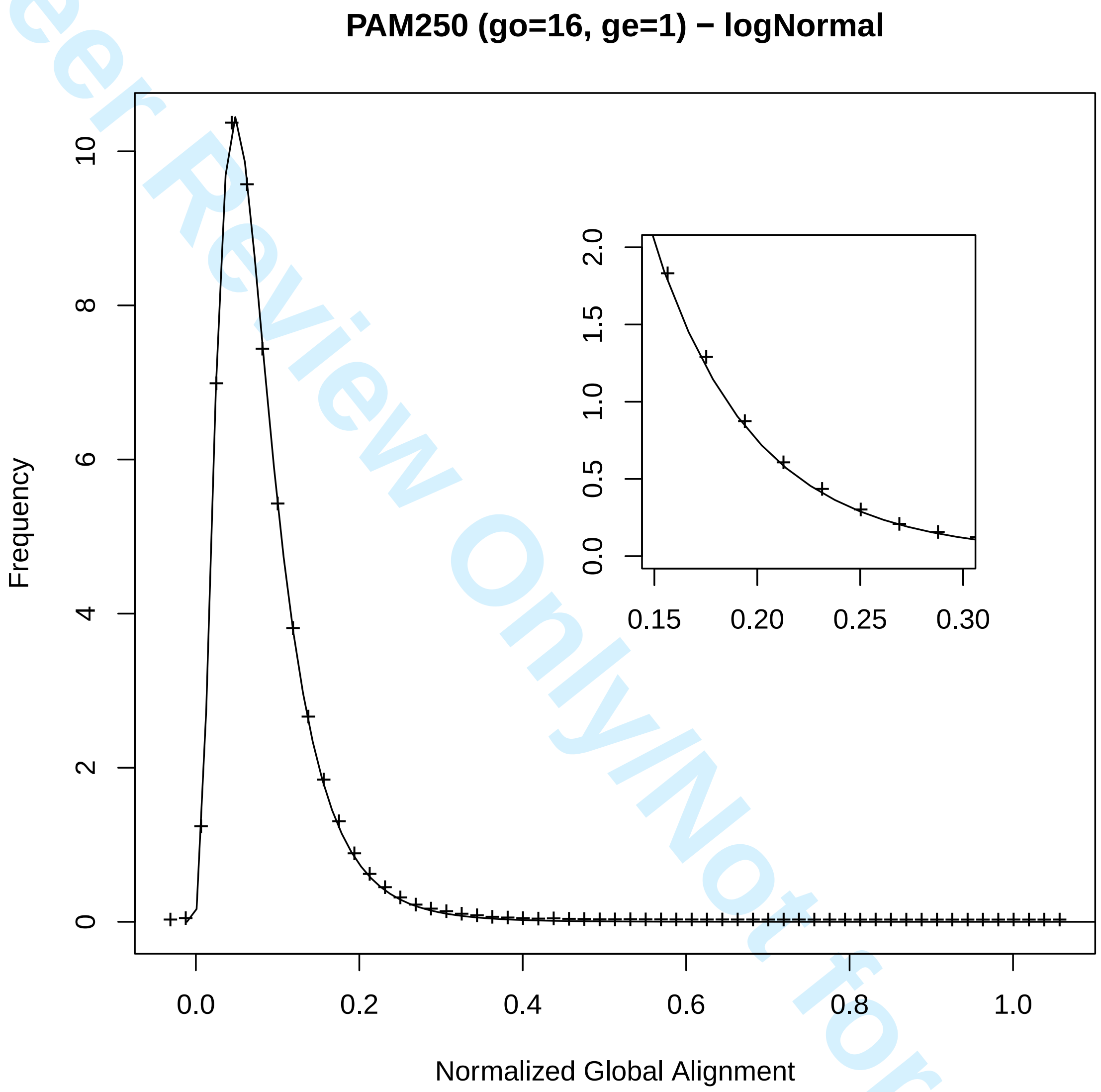

Figure 9: Distribution of NGA scores for astral40 training set, fitted to log-Normal distribution and scoring scheme PAM250, 16/1.

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 9 (of 9)

List of Tables

Table 1: log-Normal fitted parameters for NGA scores and different scoring schemes. In the last column, NGA scores needed to obtain P-values of 1×10^{-5} and 1×10^{-8} are shown.

Scoring matrix	g_o	g_e	log-Normal parameters			P-value	
			ξ	μ	σ	10^{-5}	10^{-8}
BLOSUM50	11	0	-2.0671	0.8671	0.0380	0.73	0.89
BLOSUM50	11	1	-0.0228	-1.9700	0.4422	0.90	1.65
BLOSUM50	11	2	-0.0091	-2.4850	0.5801	0.98	2.15
BLOSUM50	12	0	-0.7521	0.0263	0.0806	0.70	0.86
BLOSUM50	12	1	-0.0130	-2.2376	0.5069	0.91	1.82
BLOSUM50	12	2	-0.0069	-2.6881	0.6276	0.98	2.30
BLOSUM50	13	0	-0.3765	-0.4820	0.1228	0.67	0.85
BLOSUM50	13	1	-0.0091	-2.4604	0.5616	0.93	1.99
BLOSUM50	13	2	-0.0054	-2.8536	0.6637	0.97	2.38
BLOSUM50	14	0	-0.2112	-0.8600	0.1648	0.64	0.86
BLOSUM50	14	1	-0.0072	-2.6488	0.6068	0.93	2.12
BLOSUM50	14	2	-0.0044	-2.9820	0.6883	0.95	2.41
BLOSUM50	15	0	-0.1254	-1.1646	0.2059	0.63	0.87
BLOSUM50	15	1	-0.0059	-2.8085	0.6434	0.93	2.22
BLOSUM50	15	2	-0.0039	-3.0800	0.7046	0.92	2.39
BLOSUM50	16	0	-0.0759	-1.4251	0.2469	0.61	0.89
BLOSUM50	16	1	-0.0049	-2.9394	0.6712	0.92	2.28
BLOSUM50	16	2	-0.0035	-3.1533	0.7147	0.90	2.35
BLOSUM62	8	0	-5.8333	1.8062	0.0118	0.57	0.64
BLOSUM62	8	1	-0.0145	-2.4431	0.5014	0.72	1.78
BLOSUM62	8	2	-0.0068	-2.9841	0.6421	0.78	1.84
BLOSUM62	9	0	-0.7026	-0.0896	0.0693	0.53	0.64
BLOSUM62	9	1	-0.0088	-2.8010	0.5882	0.74	1.80
BLOSUM62	9	2	-0.0048	-3.2379	0.6917	0.75	1.73
BLOSUM62	10	0	-0.2692	-0.8072	0.1264	0.50	0.66
BLOSUM62	10	1	-0.0060	-3.0867	0.6532	0.73	1.75
BLOSUM62	10	2	-0.0040	-3.4028	0.7152	0.70	1.62
BLOSUM62	11	0	-0.1277	-1.2890	0.1827	0.47	0.68
BLOSUM62	11	1	-0.0047	-3.2929	0.6924	0.71	1.66
BLOSUM62	11	2	-0.0038	-3.5051	0.7227	0.65	1.53
BLOSUM62	12	0	-0.0652	-1.6663	0.2388	0.46	0.72
BLOSUM62	12	1	-0.0041	-3.4307	0.7112	0.67	1.57
BLOSUM62	12	2	-0.0038	-3.5679	0.7224	0.61	1.46

Continued on next page

Table 1 – continued from previous page

Scoring matrix	g_o	g_e	log-Normal parameters			P-value	
			ξ	μ	σ	10^{-5}	10^{-8}
BLOSUM62	13	0	-0.0344	-1.9789	0.2937	0.45	0.67
BLOSUM62	13	1	-0.0040	-3.5183	0.7173	0.63	1.43
BLOSUM62	13	2	-0.0039	-3.6082	0.7193	0.58	1.85
BLOSUM62	14	0	-0.0185	-2.2455	0.3464	0.45	0.65
BLOSUM62	14	1	-0.0039	-3.5756	0.7179	0.59	1.64
BLOSUM62	14	2	-0.0040	-3.6340	0.7155	0.55	1.90
PAM120	10	0	-0.5370	-0.3508	0.0787	0.45	0.56
PAM120	10	1	-0.0079	-3.4940	0.6835	0.55	1.40
PAM120	10	2	-0.0084	-3.6531	0.6665	0.44	1.08
PAM120	11	0	-0.2311	-1.0030	0.1324	0.41	0.54
PAM120	11	1	-0.0081	-3.6074	0.6749	0.47	1.19
PAM120	11	2	-0.0089	-3.6860	0.6509	0.39	0.96
PAM120	12	0	-0.1182	-1.4777	0.1869	0.39	0.53
PAM120	12	1	-0.0085	-3.6638	0.6605	0.42	1.04
PAM120	12	2	-0.0092	-3.7035	0.6387	0.37	0.88
PAM120	13	0	-0.0664	-1.8605	0.2411	0.37	0.54
PAM120	13	1	-0.0089	-3.6928	0.6472	0.38	0.93
PAM120	13	2	-0.0095	-3.7136	0.6297	0.35	0.83
PAM120	14	0	-0.0397	-2.1867	0.2948	0.36	0.55
PAM120	14	1	-0.0093	-3.7080	0.6366	0.36	0.86
PAM120	14	2	-0.0097	-3.7203	0.6236	0.34	0.79
PAM120	15	0	-0.0254	-2.4691	0.3467	0.35	0.57
PAM120	15	1	-0.0095	-3.7170	0.6287	0.35	0.82
PAM120	15	2	-0.0099	-3.7244	0.6191	0.33	0.77
PAM120	16	0	-0.0174	-2.7159	0.3962	0.34	0.59
PAM120	16	1	-0.0097	-3.7231	0.6231	0.33	0.79
PAM120	16	2	-0.0100	-3.7273	0.6159	0.32	0.75
PAM250	12	0	-1.7981	0.7209	0.0422	0.66	0.81
PAM250	12	1	-0.0410	-1.8517	0.3948	0.80	1.40
PAM250	12	2	-0.0177	-2.2641	0.5033	0.87	1.73
PAM250	13	0	-0.8419	0.0705	0.0750	0.64	0.79
PAM250	13	1	-0.0263	-2.0695	0.4458	0.82	1.51
PAM250	13	2	-0.0133	-2.4196	0.5404	0.88	1.83
PAM250	14	0	-0.4822	-0.3717	0.1085	0.61	0.79
PAM250	14	1	-0.0183	-2.2496	0.4895	0.83	1.63
PAM250	14	2	-0.0106	-2.5488	0.5708	0.88	1.91
PAM250	15	0	-0.3002	-0.7203	0.1432	0.60	0.79

Continued on next page

Table 1 – continued from previous page

Scoring matrix	g_o	g_e	log-Normal parameters			P-value	
			ξ	μ	σ	10^{-5}	10^{-8}
PAM250	15	1	-0.0137	-2.3995	0.5264	0.84	1.73
PAM250	15	2	-0.0087	-2.6550	0.5949	0.88	1.97
PAM250	16	0	-0.1960	-1.0105	0.1788	0.58	0.80
PAM250	16	1	-0.0109	-2.5245	0.5568	0.85	1.81
PAM250	16	2	-0.0073	-2.7423	0.6139	0.88	2.01

⇑

Guillermo Peris[1] and Andrés Marzal

Department de Llenguatges i Sistemes

Informátics

Universitat Jaume I, 12071, Castelló (Spain)

{peris,amarzal}@uji.es

Figure 9 (of 9)