BIOINFORMATICS

Vol. 18 no. 6 2002
Pages 802–812

Confidence measures for protein fold recognition

Ingolf Sommer, Alexander Zien, Niklas von Öhsen, Ralf Zimmer and Thomas Lengauer*

Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany

Received on June 26, 2001; revised on November 14, 2001, December 20, 2001; accepted on January 7, 2002

ABSTRACT

Motivation: We present an extensive evaluation of different methods and criteria to detect remote homologs of a given protein sequence. We investigate two associated problems: first, to develop a sensitive searching method to identify possible candidates and, second, to assign a confidence to the putative candidates in order to select the best one.

For searching methods where the score distributions are known, p-values are used as confidence measure with great success. For the cases where such theoretical backing is absent, we propose empirical approximations to p-values for searching procedures.

Results: As a baseline, we review the performances of different methods for detecting remote protein folds (sequence alignment and threading, with and without sequence profiles, global and local). The analysis is performed on a large representative set of protein structures.

For fold recognition, we find that methods using sequence profiles generally perform better than methods using plain sequences, and that threading methods perform better than sequence alignment methods.

In order to assess the quality of the predictions made, we establish and compare several confidence measures, including raw scores, z-scores, raw score gaps, z-score gaps, and different methods of p-value estimation. We work our way from the theoretically well backed local scores towards more explorative global and threading scores.

The methods for assessing the statistical significance of predictions are compared using specificity–sensitivity plots. For local alignment techniques we find that p-value methods work best, albeit computationally cheaper methods such as those based on score gaps achieve similar performance. For global methods where no theory is available methods based on score gaps work best.

By using the score gap functions as the measure of confidence we improve the more powerful fold recognition methods for which p-values are unavailable.

Availability: The benchmark set is available upon request.

Contact: ingolf.sommer@gmd.de

*To whom correspondence should be addressed.

INTRODUCTION

Newly discovered amino acid sequences for proteins whose structure and function are unknown so far are commonly searched against databases of proteins that have been studied already in order to find similarities in structure and function. The comparison method for this search can be sequence–sequence or sequence–structure based. In order to compare, an alignment is computed of the target protein sequence (whose structure we are searching) with a template protein (whose structure we know). For a sequence–sequence alignment, the alignment algorithm optimizes a certain scoring function that quantifies the similarities of the amino acids at individual positions. For a sequence–structure alignment, also known as threading, the scoring function is designed to capture the essence of structural similarity among proteins.

These scores are supposed to be comparable between different proteins, since we want to select as candidate for our structural model the template that achieves the highest alignment score to the target protein. The currently available scoring functions are inaccurate, however. Thus, it is very helpful if the method can augment the generated alignment and its score with a statistical significance value which captures the confidence that we can put into the prediction of the structural model.

While important for protein alignment and threading as stand-alone tools, significance scores are even more essential if protein alignment is used in an automated cascade of tools for protein structure prediction (cf. Figure 1).

In this paper we analyse the performance of several variants of the 123D protein threading method (Alexandrov *et al.*, 1996) and compare it to several variants of

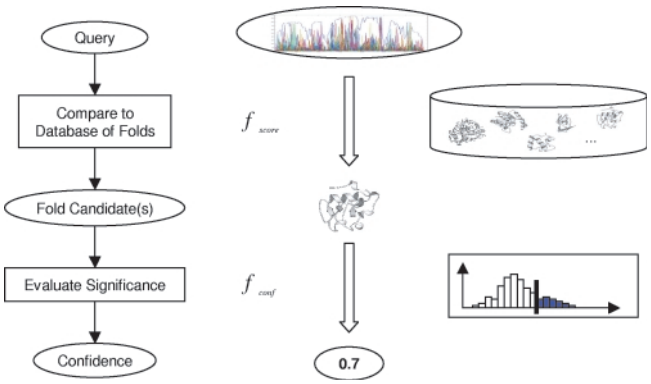


Fig. 1. Overview of scoring and confidence measure estimation: a sequence profile or sequence is scored versus a database of folds and then a confidence measure is computed for the predicted fold candidate.

optimal sequence alignment. Where theoretically available, we analyse the statistical significance of scores. For the other methods, we propose empirical approximations to p-values and other confidence measures and evaluate their validity.

EXPERIMENTAL SETUP

Protein data

In order to estimate score distributions and to evaluate different confidence measures we aligned and threaded a set of amino acid target sequences versus a reference set of structures of protein domains. This application scenario is typical for identifying domains which could serve as structural models for (parts of) the sequence. This setup was designed to resemble experiments with unknown sequences like for the Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Benner *et al.*, 1992; Barton and Russell, 1993; Defay and Cohen, 1995; Jones, 1997; Fischer *et al.*, 1999, 2000).

The protein domains were chosen as representative for the protein domains deposited in the Protein Data Bank (PDB)[†](Berman *et al.*, 2000) and classified according to the Structural Classification of Proteins database (SCOP)[‡](Murzin *et al.*, 1995). The domains with a maximum of 40% sequence identity were used as computed by the ASTRAL-Server (Brenner *et al.*, 2000). We refer to this set of domains as PDB40D.

As target sequences we chose sequences of proteins which are deposited in the PDB. We refer to the set of target sequences that we used for our experiments as PDB40C. The set contains each protein chain with at least one domain belonging to the PDB40D. Only those proteins were included of which at least all C_α atom coordinates are deposited in the PDB. Proteins with domains consisting of more than one chain were not included, since they do not allow for an unambiguous evaluation of the results.

The PDB40D contains 2860 protein domains, consisting of 22 to 925 amino acids. The average number of amino acids per domain is 182. The PDB40C set contains 2232 proteins, 1666 of which are single-domain and 566 of which are multi-domain proteins. The length of the sequences ranges from 22 to 1296, with a mean of 244 amino acids.

The knowledge of the structures of the target sequences allows for evaluating the threading results. On the other hand, admittedly, selecting a set of sequences on the basis of known structures biases the selection towards proteins the structures of which have been determined experimentally.

Related and unrelated pairs. We classify the protein pairs in PDB40C × PDB40D, that are aligned against each other, into related (positive) pairs or unrelated (negative) pairs, according to the SCOP classification. Each protein chain contains one or more domains with potentially different folds. If one of these folds is the same as the fold of the structure compared against, the pair is classified as related. It is classified as unrelated otherwise.

Alignment and threading parameters

We compare eight different methods for detecting remotely homologous protein folds, resulting from three independent binary choices: sequence alignment versus threading $[s/t]$, with plain sequences (i.e. without frequency profiles) versus with frequency profiles $[s/f]$, and local versus global $[l/g]$. We denote the resulting methods with ssl, sfl, tsl, tfl, ssg, sfg, tsg, tfg.

Optimal global sequence alignment with affine gap costs is done with a standard Gotoh algorithm (Gotoh, 1982), local sequence alignment is performed using the Smith–Waterman algorithm (Smith and Waterman, 1981). The alignment algorithms were applied to plain sequences as well as to frequency profiles. We refer to frequency profiles as a specific method of implementing sequence profiles as described below. The frequency profiles were generated and used on the target-sequence side only (scoring with frequency profiles on both sides was analysed separately, see (von Öhsen and Zimmer, 2001)). We use 12 as gap insertion and 2 as gap extension costs, respectively, for the sequence alignment.

The basic version of the 123D threading tool is described in (Alexandrov *et al.*, 1996). All threading experiments described in this paper were performed

[†] Version of February 29, 2000
[‡] Release 1.50 of February 29, 2000, classifying 10650 PDB protein entries and 24186 protein domains

with a new implementation that represents an advanced state of developoment in this tool in two ways: (1) Instead of threading the target sequence itself against a template structure with its native amino acids, we can also employ sequence profiles. (2) The scoring function, originally a sum of inverse Boltzmann derived potentials, is tuned by optimally weighting the individual contributions against each other. This was shown to improve performance for a broad variety of application scenarios, including different protocols and threading with profiles (Zien *et al.*, 2000). This parameter calibration also supplied values for gap insertion and extension costs, which we used for this study. In the following we refer to global threading as optimized with the 123D program using these parameters. With O(sequence length $*$ structure length) its runtime complexity is of the same order as that for sequence alignment.

The amino acid frequency profiles for the sequences (Gribskov *et al.*, 1987; Park *et al.*, 1998) were computed as follows: First, we generated a multiple alignment by running Psi-Blast (Altschul *et al.*, 1997) against the KIND (Kallberg and Persson, 1999) database. Then, for each sequence s_i in the alignment a weight w_i was computed that is supposed to compensate for overrepresentation of similar protein hits. The sequence weighting algorithm extends the idea presented in (Henikoff and Henikoff, 1994). Henikoff's algorithm was shown to approximately maximize the entropy of the resulting profile (Krogh and Mitchison, 1995). In contrast, we distribute the total weight per column over the amino acids according to their relative background frequencies. The profile was generated by simply computing positional relative frequencies based on the weighted sequences, augmented by a small number of pseudo-counts (we used 0.1). The resulting profile approximates one that has minimal relative entropy regarding the amino acid background distribution.

FOLD RECOGNITION AND CONFIDENCE MEASURE METHODS

In the fold recognition protocol, for each target protein sequence we try to find a template protein structure, thereby identifying a corresponding fold class. This is done by evaluating a scoring function for each target–template pair and sorting the template scores for one target. Template domains from the target sequence are excluded from performance evaluation (leave-one-out protocol). The fold of the highest scoring template is then predicted to be the fold of the target sequence. For this template a confidence score can be computed in order to estimate the validity of the prediction.

Some confidence measures, like e.g. the p-values, can be computed only if the score distribution is known. However, this is not the case for all alignment methods considered.

Statistical significance of alignment and threading scores

For optimal local gapless sequence alignments of independent random sequences the scores are known to be asymptotically extreme-value or Gumbel distributed (Karlin and Altschul, 1990):

$$P(\text{score} > t) \approx 1 - \mathrm{e}^{-K\mathrm{e}^{-\lambda t}}$$

where λ depends only on the scoring system and K depends on the scoring system and the sequence lengths, such that the distribution reflects the fact that the chance of spurious high scores increases with sequence lengths. The dependence of the parameters on the scoring system and sequence lengths is known (Karlin and Altschul, 1990).

For local alignments with gaps of unrelated biological sequences no theory is available, however there is a lot of evidence that the distribution is still of extreme-value form and parameters can be fitted experimentally (Waterman and Vingron, 1994; Altschul and Gish, 1996; Pearson, 1998; Levitt and Gerstein, 1998; Mott, 2000). Local alignments with sequence-profiles were also shown to follow an extreme-value distribution (Mott, 2000).

For optimal global alignments, whether with plain sequences or sequence profiles, neither the family of distributions nor the dependence of the expected score (or of other parameters) on the sequence lengths is known, to the best of our knowledge.

The situation is similar for 123D threading: The local threading scores of sample sequence-structure pairs closely follow a Gumbel distribution (an example is shown in Figure 2). For global threading the distribution is unknown. Sample score distributions depend on the length of the sequences but neither resemble a Gaussian nor a Gumbel distribution (examples are shown in Figures 3 and 4).

Scoring and confidence functions

Let SEQ be a set of amino acid sequences of proteins and STR be a set of structures of proteins, then a scoring function

$$\begin{aligned} f_{\text{score}} : \text{SEQ} \times \text{STR} &\longrightarrow \mathbb{R} \\ (\text{seq}, \text{str}) &\longmapsto f_{\text{score}}(\text{seq}, \text{str}) \end{aligned}$$

is applied to each target sequence seq to find a related template structure pred_{seq} with

$$f_{\text{score}}(\text{seq}, \text{pred}_{\text{seq}}) = \max_{\text{str} \in \text{STR}_{\text{seq}}} f_{\text{score}}(\text{seq}, \text{str})$$

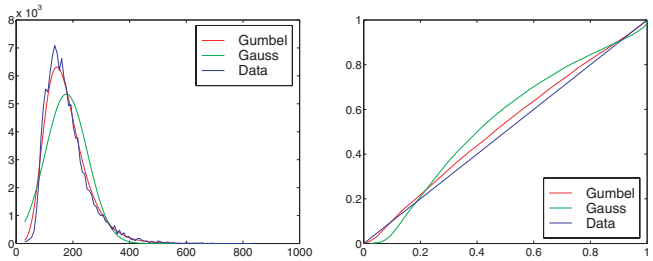


Fig. 2. Gauss and Gumbel functions fitted to local threading scores of 10585 representatively selected, unrelated sequence-structure pairs with a sequence length between 70 and 80 amino acids. Left: density functions (x-axis: score, y-axis: density), right: quantile–quantile plots (x-axis: distribution quantile as listed in the legend, y-axis: data quantile); fits were performed as described in appendix A.

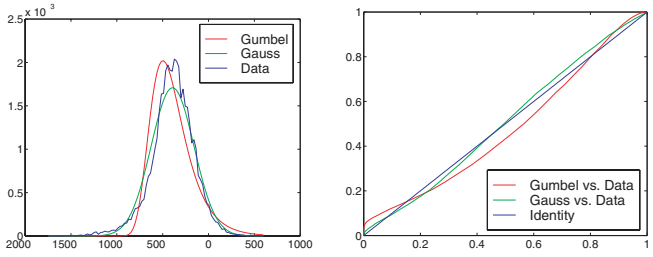


Fig. 3. Gauss and Gumbel functions fitted to global threading scores of 10585 representatively selected, unrelated sequence-structure pairs with a sequence length between 70 and 80 amino acids. Left: density functions (x-axis: score, y-axis: density), right: quantile–quantile plots (x-axis: distribution quantile as listed in the legend, y-axis: data quantile); fits were performed as described in appendix A. Note that, in contrast to the situation depicted in Figure 2, where the Gumbel distribution describes the data more accurately and where the dependence of the parameters of the distribution of the sequence and structure lengths is known, the Gaussian distribution suits the data better. The dependence of the Gaussian distribution of the sequence lengths is unknown. We tested the complete data set used in the experiments for both kinds of distributions: The data were separated into clusters of similar sequence and structure lengths, and for each cluster a χ^2-test versus a fitted Gaussian distribution and a fitted Gumbel distribution was performed. For all clusters with enough data points (those with lengths smaller than 600 amino acids) similarity was rejected at a confidence level of 0.99.

where we omit scoring the target sequence versus itself by defining

$$\mathrm{STR}_{\mathrm{seq}} = \{\mathrm{str} \in \mathrm{STR} | \mathrm{sequence(str)} \text{ not subsequence of } \mathrm{seq}\}.$$

The fold of the structure $\mathrm{pred}_{\mathrm{seq}}$ is predicted to be the most plausible fold for sequence seq. A confidence function

$$\begin{aligned} f_{\mathrm{conf}} : \mathrm{SEQ} \times \mathrm{STR} &\longrightarrow \mathbb{R} \\ (\mathrm{seq}, \mathrm{str}) &\longmapsto f_{\mathrm{conf}}(\mathrm{seq}, \mathrm{str}) \end{aligned}$$

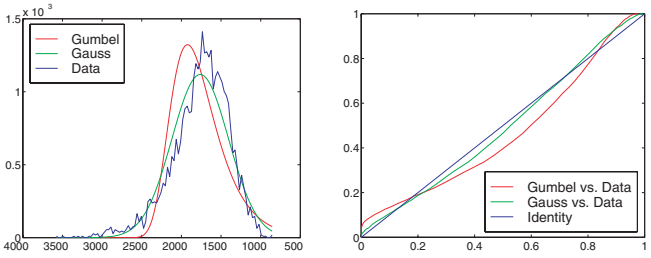


Fig. 4. Gauss and Gumbel functions fitted to global threading scores of 3996 representatively selected, unrelated sequence-structure pairs with a sequence length between 70 and 80, structure length between 270 and 280 amino acids. Left: density functions (x-axis: score, y-axis: density), right: quantile–quantile plots (x-axis: distribution quantile as listed in the legend, y-axis: data quantile); fits were performed as described in appendix A. Similar to the situation in Figure 3, the Gaussian distribution approximates the data better.

is then applied to quantitatively estimate the validity of the prediction as $f_{\mathrm{conf}}(\mathrm{seq}, \mathrm{pred}_{\mathrm{seq}})$. This measure tells how much trust to put into the result of the alignment or threading, i.e. how certain we are that the prediction is correct.

The confidence function f_{conf} is usually calibrated over a set of sequences and structures $\mathrm{CAL} \subset \mathrm{SEQ} \times \mathrm{STR}$. For example, z-scores (see section below) are computed by estimating mean and standard deviation of the distribution of scores for $\mathrm{CAL} = \mathrm{seq} \times \mathrm{STR}$ and normalizing scores according to these. Metainformation, like knowledge of the relatedness of sequences and structures or of the relatedness of different structures, may also be used for calibration.

Apart from the calibration, scoring functions and confidence functions require the same input and output and can be used interchangeably. Next, we present the scoring and confidence functions that we compared in our experiments.

Raw score function (s)

With raw scores we denote the score of an optimal threading or sequence alignment. Thus, for plain sequences, $s_{\mathrm{ssg}}(\mathrm{seq}, \mathrm{str})$ is the score output of the Gotoh alignment of seq and str, while s_{ssl} denotes the score of the Smith-Waterman alignment, s_{tsl} is the score computed with local 123D threading, and s_{tsg} is the score computed with global 123D threading. Computed with frequency profiles, we denote the raw scores with $s_{\mathrm{f.}}$ accordingly.

Trivially, the raw scores can be used as scoring functions $f_{\mathrm{score}} = s$ and also as confidence measures by $f_{\mathrm{conf}} = s$. The higher the score, the more we trust in the result. Obviously this disregards the influence of sequence and structure lengths on the score. Furthermore, raw scores computed with different methods (e.g. threading, alignment)

or parameter sets (e.g. different gap costs) are in general incomparable.

z-Score function (z)

A common way of rendering raw scores comparable across sequences is to compute z-scores, by using the set of all template proteins that the target sequence is threaded or aligned against to normalize the threading scores. Mean $\hat{\mu}_{\text{seq}}$ and standard deviation $\hat{\sigma}_{\text{seq}}$ are estimated from the scores of one target sequence seq versus its set of templates STR_{seq} as

$$\hat{\mu}_{\text{seq}} = \frac{1}{|\text{STR}_{\text{seq}}|} \sum_{\text{str}\in\text{STR}_{\text{seq}}} s(\text{seq}, \text{str})$$

$$\hat{\sigma}^2_{\text{seq}} = \frac{1}{|\text{STR}_{\text{seq}} - 1|} \sum_{\text{str}\in\text{STR}_{\text{seq}}} \big(s(\text{seq}, \text{str}) - \hat{\mu}_{\text{seq}}\big)^2.$$

The scores are then normalized into z-scores using

$$z(\text{seq}, \text{str}) = \frac{s(\text{seq}, \text{str}) - \hat{\mu}_{\text{seq}}}{\hat{\sigma}_{\text{seq}}}.$$

Fitted p-values (*a*,*b*)

With a known score distribution, the probability that an alignment score of at least this magnitude occurs by chance can be computed. This probability is generally called *p-value*. The lower the p-value, the more certain we are that the two sequences are related. As noted previously, the parametric form of the score distribution is known for local methods only.

The expected score of an optimal local alignment is known to depend on the lengths of the two aligned sequences. For gapless alignments, this dependency is shown to be either logarithmic ($E(\text{score}) = c\log(l_1 l_2)$) or linear ($E(\text{score}) = c\sqrt{l_1 l_2}$) for scoring matrices with negative or positive expectation value, respectively, with a transition phase in between (Arratia and Waterman, 1994).

We fit an extreme value distribution parametrized by

$$P(\text{score} > t) = 1 - \mathrm{e}^{-\mathrm{e}^{-f(t,l)}}$$

where

$$f(t, l) = \frac{\pi}{6}\left(\frac{t - (\beta l + \alpha)}{\nu l + \mu}\right) + \gamma$$

and l is an appropriate function of the lengths l_1, l_2 of the aligned sequences, i.e. $l = \log(l_1 l_2)$ or $l = \sqrt{l_1 l_2}$. The Eulerian number γ is defined in Appendix A. We determine $(\alpha, \beta, \mu, \nu)$ empirically by fitting the function to the data in the following way. First, α and β are estimated by performing a linear regression of the observed values score against the corresponding values l. In order to make the procedure more robust, we subsequently mark as outliers all pairs with s being more than -3 or $+5$ standard deviations away from the value expected from the fit, $\beta l + \alpha$. We repeat the linear regression until the set of outliers remains unchanged. In the second step, we determine μ and ν by a simple linear regression of the estimated standard deviation of the scores with respect to the corresponding expected scores, $|\text{score} - (\beta l + \alpha)|$, against the length values l.

For alignments of both plain sequences and frequency profiles against sequences, the length dependence of the scores seems to be in the logarithmic phase. For threading, the dependency is modeled more accurately by the linear model. This seems to be inconsistent with the observation that the naive expectation values of all employed potentials are negative. However, the 123D threading potentials depend on the secondary structure of the aligned position in the structure. While neighboring amino acids are often considered independent, this assumption is obviously violated for secondary structure elements, which tend to occur in consecutive stretches. Thus, a fundamental prerequisite of the derivation of the Gumbel distribution is violated in this case. However, there is empirical evidence that the scores still approximately follow a Gumbel distribution with linear length dependence. Our results (see below) show that p-values computed according to this model are meaningful, to some extent.

We consider two different protocols for parameter fitting: In the first case (*a*), the fit is performed separately for each sequence: CAL = seq × STR. In the other case (*b*), the data arising from all sequences are joined and the fit is carried out over all sequence–structure pairs: CAL = SEQ × STR.

Tabulated p-values (*u*)

The global threading and alignment scores are known to depend on the sequence and structure length, but the nature of that dependence is not known. An approach to estimating score distributions from data is to generate tables of score-percentiles from a set of unrelated sequence-structure pairs. Later, these tables are used to look up the estimated probability of a new score to belong to this set of unrelated pairs.

Cross-validation. For estimating the distributions, i.e. to calibrate the confidence functions, we work with a fivefold cross-validation. The PDB40C is separated into five subsets of equal size that do not have overlaps in SCOP-families (i.e. we assert that there is no SCOP family with members belonging to more than one test set). Proteins of four of the five subsets are threaded against all members of the PDB40D, and the unrelated protein-pairs are used to estimate the score distribution, which is then evaluated for the scores of the proteins in the fifth subset threaded versus the members of the PDB40D.

Length dependence. The scores s of the unrelated pairs of the calibration set are stored in tables according to lengths of sequences and structures. One table is generated for each parameter setting ssl, sfl,..., tfg. For this, bins are defined by b_{seq} intervals for the sequence lengths and b_{str} intervals for the structure lengths. The intervals by sequence length are chosen such that each interval contains the same number of sequences, and analogously for the intervals by structure length. Note that this process of binning does not imply an equal number of sequence-structure pairs for each bin. For the data set we used, with $b_{\text{seq}} = 10$ and $b_{\text{str}} = 10$, the resulting interval boundaries occur at lengths [0, 70, 100, 126, 160, 204, 247, 302, 367, 476, 1296, ∞) for the sequences, and [0, 63, 87, 104, 123, 145, 176, 218, 269, 348, 925, ∞) for the structures.

To estimate the p-value

$$P(\text{score} > t | \text{sequence length, structure length})$$

for a new score t with given sequence length and structure length, the previously generated table which suits that sequence and structure length is searched, and the relative frequency of scores above the threshold t is used as an estimate P_{est} for the p-value. Thus

$$_u(\text{seq, str}) = P_{\text{est}}(\text{score} > {}_s(\text{seq, str}) | {}_{\text{length}}(\text{seq}), {}_{\text{length}}(\text{str})).$$

Since this leaves all examples with a score higher than the highest score seen during calibration with a p-value of zero, for later comparisons we sort all these pairs according to their raw scores.

Raw score gaps (*sg*)

For a target sequence seq, the *raw score gap* is the difference $sg(\text{seq, str}) = s(\text{seq, str}) - s(\text{seq, next(str)})$ of the raw score of a template protein str and the next best raw score of a template protein belonging to a different fold, next(str). This gap can be computed for all but the lowest scoring fold. Thus, it is always defined for the highest scoring fold and can be used as a confidence measure for fold recognition. Intuitively, the larger the difference (score gap) between the best fold and the highest scoring alternative, the more confidence we have in the prediction. The same limitation that applies to the raw scores applies here: score gaps computed with different parameter sets are incomparable.

The raw score gap is reminiscent of a frequently used model selection criterion, the log likelihood ratio of two hypotheses. To see this, we need to recall that the raw scores are the sum of substitution scores and insertion and deletion costs. At each alignment position, the substitution score is the log likelihood ratio of an evolutionary amino acid substitution against an independent pairing according to the background frequencies (Altschul, 1991). Therefore, the positional score gap corresponds to the log likelihood ratio of the evolutionary change from the predicted fold as compared to the change from the next highest ranking fold. Since often only one fold class is correct, the score gap should be large for many true predictions. Since wrong folds should be similarly unlikely, the score gap should be small if a wrong fold ranks highest. The encouraging results (see below) seem to back this reasoning.

z-Score gaps (*zg*)

Analogously, the *z-score gap* can be computed for each target sequence and used as the confidence measure. It is the difference of the z-score of a template protein and the next best z-score of a template protein belonging to a different fold $zg(\text{seq, str}) = z(\text{seq, str}) - z(\text{seq, next(str)})$.

RESULTS AND DISCUSSION

With our experiments we address two questions: (I) Which scoring function $f_{\text{score}} \in \{s, z, a, b, u\}$ yields the best fold recognition performance? (II) Which confidence function $f_{\text{conf}} \in \{s, z, a, b, u, sg, zg\}$ is best to evaluate the prediction produced with a given scoring function?

This corresponds to evaluating the first and the second step of the fold recognition cascade as depicted in Figure 1 successively. The performance values computed in the first step serve as a baseline for evaluating the confidence functions later on.

(I) Fold recognition performance

In order to evaluate the fold recognition performance of the different methods, we consider two test sets. Our reference is the test set PDB40C $\times$ PDB40D as described in a previous section. This complete test set is also used for benchmarking the confidence measures. The other test set is a subset containing the more difficult cases.

Complete test set. For 1581 (70.1%) of the proteins in the PDB40C there is a remote member of the same SCOP family in the PDB40D, for 363 (16.3%) of the proteins in the PDB40C there is a member of the same SCOP superfamily in the PDB40D. For 121 (5.4%) there is a structure with corresponding fold, but none with a corresponding superfamily, and for 167 (7.5%) there is no corresponding fold (other than themselves) in the PDB40D.

When we test fold recognition on this set, the maximal theoretical performance would be 92.5%, since 7.5% of the sequences have no corresponding fold in the database. We included those sequences into the test set, in order to simulate the 'real world' situation, where a new protein does not necessarily resemble a known structure.

The fold recognition performances of the different

Table 1. Fold recognition performance results on the PDB40C × PDB40D for local (above) and global (below) methods. Performances are given in percent, the maximum of each column is set in boldface. The scoring functions are listed in the rows, the threading and alignment parameters in the columns of the table. As a comparative value, using Psi-Blast (ten iterations with default parameters) in the same protocoll results in a performance ratio of 66.8%. To estimate the reliability of the observed performance rates, we calculate pessimistic approximations to their standard deviations. We do so by modeling the number of successful predictions as a binomial distribution $\mathcal{B}(n; p)$, where n is the size of the benchmark set and p approximately equals the observed success rate. The standard deviation of this number is $\sqrt{np(1-p)}$, which corresponds to about 1% of 2232.

	ssl	sfl	tsl	tfl
s	61.4	69.6	61.6	**70.7**
z	61.4	69.6	61.6	**70.7**
a	52.4	64.8	60.2	66.2
b	**63.5**	**69.7**	**65.4**	**70.7**
u	62.9	68.8	63.2	69.8
	ssg	sfg	tsg	tfg
s	57.2	66.8	60.6	**72.7**
z	57.2	66.8	60.6	**72.7**
u	**59.2**	**67.3**	**62.3**	70.2

Table 2. Fold recognition performance results on 121 difficult sequences of the PDB40C × PDB40D test set (see text) for local (above) and global (below) methods. Performances are given in percent, the maximum of each column is in boldface. The fold recognition performance with Psi-Blast is 6.6%. For detecting proteins with similar structure but remote homology, sequence based methods are not necessarily adequate. However, they perform better than simple guessing.

	ssl	sfl	tsl	tfl
s	5.8	**9.1**	9.1	**14.9**
z	5.8	**9.1**	9.1	**14.9**
a	3.3	5.8	6.6	8.3
b	**8.3**	**9.1**	**12.4**	14.1
u	5.0	**9.1**	9.1	**14.9**
	ssg	sfg	tsg	tfg
s	9.1	**11.6**	**10.7**	**19.8**
z	9.1	**11.6**	**10.7**	**19.8**
u	**9.9**	9.9	9.1	16.5

scoring functions (raw score (s), z-score (z), estimated p-values (a, b) and tabulated p-values (u)) are listed in Table 1 for the different methods sequence-alignment (s..) and threading (t..), with sequences (.s.) and frequency profiles (.f.) for local (..l) and global (..g) alignments. The recognition rates range from 51.9% for a_{ssl} to 72.7% for s_{tfg} with most rates between 60–70%.

Overall, threading performs better than plain sequence alignment, and both alignment and threading perform better with frequency profiles than without. This makes threading with frequency profiles the method of choice for predictions.

Test set reduced to difficult cases. In order to study the performance differences of alignment and threading in more detail, we have chosen a subset of the test set above with 121 more difficult sequences. These sequences have a corresponding fold, but no corresponding superfamily within the PDB40D. The corresponding fold recognition performances are listed in Table 2.

Consistently with the complete test set, alignment and threading perform better with frequency profiles than without, and threading performs better than plain sequence alignment. The advantage of the threading method is clearly visible here, performance rates are more than doubled, from plain sequence alignment to threading with frequency profiles (5.8–14.9% for local, and 9.1–19.8% for global modes).

Scoring function performance. For the local methods, the distribution of the raw scores is known and the p-value fitting score functions (a,b) make use of this knowledge; when enough data are used to fit the parameters of the scoring functions (b), these scoring functions work best. The performance rates of the tabulation method (u), which does not make any assumption about the score distribution, are not drastically below the fitted p-values (a,b).

(II) Comparison of confidence measures

On the test set PDB40C × PDB40D and the fold recognition results obtained with the scoring function $f_{\mathrm{score}} = s$, for the different parameter sets we compare the confidence functions $f_{\mathrm{conf}} \in \{s, z, a, b, u, sg, zg\}$.

Evaluation criteria. For each confidence function, the predictions $\{(\mathrm{seq}, \mathrm{pred}_{\mathrm{seq}}) | \mathrm{seq} \in \mathrm{PDB40C}\}$ are sorted according to their confidence scores $f_{\mathrm{conf}}(\mathrm{seq}, \mathrm{pred}_{\mathrm{seq}})$. We thus evaluate the recognition performance for the case that a fold is predicted only for those sequences which attain a confidence higher than a given threshold (i.e. those that the confidence measure declares as 'reliable' predictions). For each of the methods and for each threshold t we count the number of predictions,

$$\mathrm{predicted}(t) = \left|\{\mathrm{seq} | f_{\mathrm{conf}}(\mathrm{seq}, \mathrm{pred}_{\mathrm{seq}}) > t\}\right|$$

and the number of true positives, false positives, true negatives, and false negatives

$$\begin{aligned} tp(t) &= \big|\{\text{seq}|(f_{\text{conf}}(\text{seq}, \text{pred}_{\text{seq}}) > t) \\ &\qquad \wedge \quad \text{rel}(\text{seq}, \text{pred}_{\text{seq}})\}\big| \\ fp(t) &= \big|\{\text{seq}|(f_{\text{conf}}(\text{seq}, \text{pred}_{\text{seq}}) > t) \\ &\qquad \wedge \neg\text{rel}(\text{seq}, \text{pred}_{\text{seq}})\}\big| \\ tn(t) &= \big|\{\text{seq}|(f_{\text{conf}}(\text{seq}, \text{pred}_{\text{seq}}) \leq t) \\ &\qquad \wedge \neg\text{rel}(\text{seq}, \text{pred}_{\text{seq}})\}\big| \\ fn(t) &= \big|\{\text{seq}|(f_{\text{conf}}(\text{seq}, \text{pred}_{\text{seq}}) \leq t) \\ &\qquad \wedge \quad \text{rel}(\text{seq}, \text{pred}_{\text{seq}})\}\big| \end{aligned}$$

with

$$\text{rel}(\text{seq}, \text{pred}_{\text{seq}}) = (\text{fold}(\text{pred}_{\text{seq}}) \cap \text{fold}(\text{seq}) \neq \emptyset)$$

where fold(seq) is the set of all SCOP folds into which the protein with sequence seq is categorized, and fold(str) is the set of all SCOP folds into which the protein with structure str is categorized[§]. The number of related pairs remains constant throughout

$$related = \big|\{\text{seq}|\text{rel}(\text{seq}, \text{pred}_{\text{seq}})\}\big|.$$

For each threshold t, specificity, the estimated probability that a prediction made is correct, and sensitivity, the estimated probability of predicting when the prediction is correct, are defined as (Bailey *et al.*, 2000)

$$\begin{aligned} \text{specificity}(t) &= \frac{tp(t)}{tp(t) + fp(t)} = \frac{tp(t)}{\text{predicted}(t)} \\ \text{sensitivity}(t) &= \frac{tp(t)}{tp(t) + fn(t)} = \frac{tp(t)}{\text{related}}. \end{aligned}$$

The probability that a prediction with a confidence at least as large as t is incorrect can be estimated with $\frac{fp(t)}{\text{predicted}(t)}$.

For the parameter sets ssl, sfl, tsl, tfl, ssg, sfg, tsg, and tfg, for each confidence function $f_{\text{conf}} \in \{s, z, a, b, u, sg, zg\}$, specificity–sensitivity plots are shown in Figures 5 and 6.

DISCUSSION

In the plots specificity is shown on the x-axis, sensitivity on the y-axis. By definition the lines for all methods join at the endpoints. The specificity in the upper left endpoint corresponds to the fold-recognition rate, as listed in Table 1. These endpoints thus show the performance of the underlying scoring function. A criterion for interpreting the goodness of a confidence measure is the form of the corresponding curve; the closer the curve gets to the upper right corner, the better.

While the tendency regarding fold-recognition, as pointed out before, is clearly visible—threading performs better than sequence alignment, and methods with frequency profiles outperform methods without—, the interpretation of the different confidence measures is more subtle: For all the local methods, the tabulated p-values (u) perform nearly as well as the p-values (b) calculated according to the Karlin–Altschul model with parameters estimated on the same data basis. We therefore conjecture that these tabulated p-values are reasonable estimates for the significances of global alignments, where no theoretical model is available.

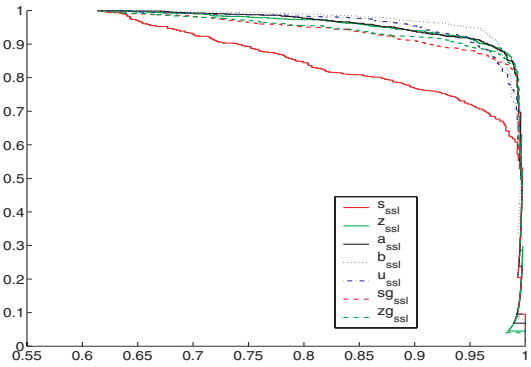
Further we notice that the p-values (u or b), tabulated and estimated over all pairs, compared to the p-values (a), with parameters fitted on a per sequence basis, have a significant advantage for the plain-sequence cases (.s.), but not for frequency profile cases (.f.). In the latter cases, to the best of our understanding, we are missing an additional, unknown parameter. This might be the entropy of the frequency profile, or the number of homologous sequences that the profile was generated from. For the per sequence fit (a), this parameter remains constant over all data and can thus be compensated for by the fitting to the remaining parameters.

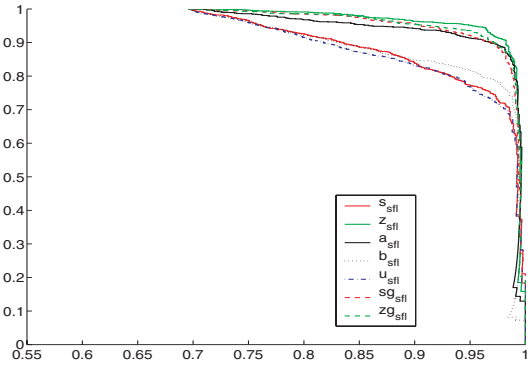
For the local parameter sets (..l), z-scores (z), score gaps (sg), z-score gaps (zg) and p-values fitted per sequence (a) perform better than the raw scores (s). As mentioned above, the p-values, tabulated (u) and estimated (b) for all pairs, perform competitively for the plain sequence methods (.s.) only.

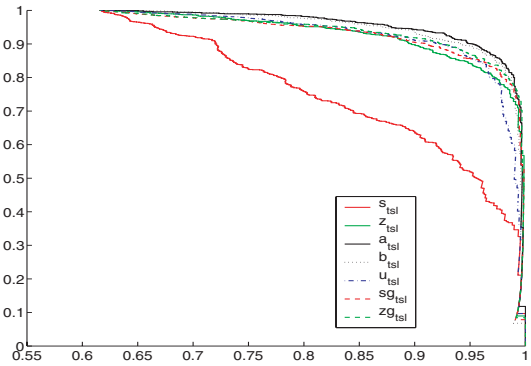
For the global methods (..g), score gaps (sg), z-score gaps (zg) and tabulated p-values (u) perform significantly better than raw scores (s) and z-scores (z); again, the p-values (u) suffer in the frequency profile cases (.fg). Clearly, the z-scores are not adequate for global methods (note the drop of the z-score in the (..g) plots). Contrary to local methods, global methods can produce negative scores. These scores can achieve a high z-score being much better than average for still very negative scores belonging to unrelated candidate pairs (we found examples for this in our data).

Albeit computationally much simpler to handle than tabulated p-values, we find that the score gaps (sg and zg) perform highly competitive as confidence measures for all methods proposed. Unfortunately, these score gaps do not compare directly for different parameter sets, as p-values do; thus the disadvantage that remains is the need to calibrate the score gaps with example data.

[§] A multidomain protein can belong to more than one fold class; as an example consider the two domain protein with PDB key 1amy; sequence(1*amy*) = *QVLFQGFNWES...AVWEKI*, fold(sequence(1*amy*)) = { 'alpha-Amylases, C-terminal beta-sheet domain', 'TIM beta/alpha-barrel' }.







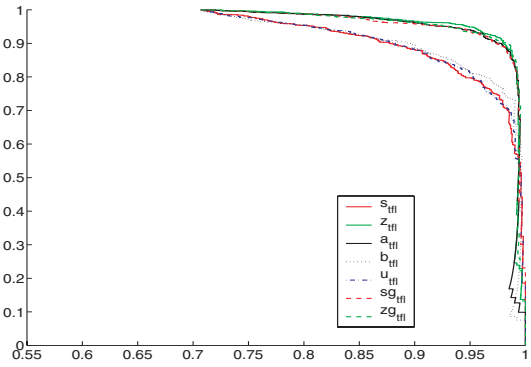
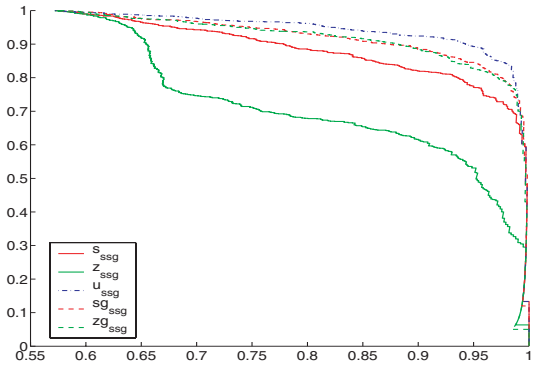
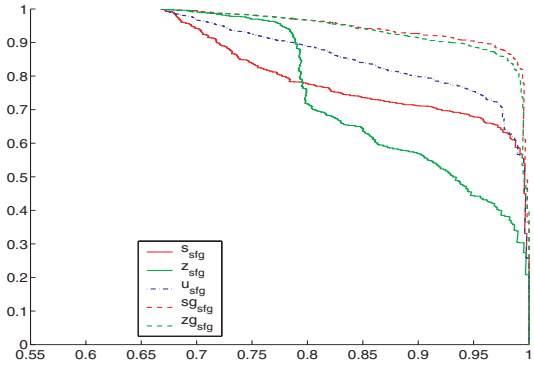
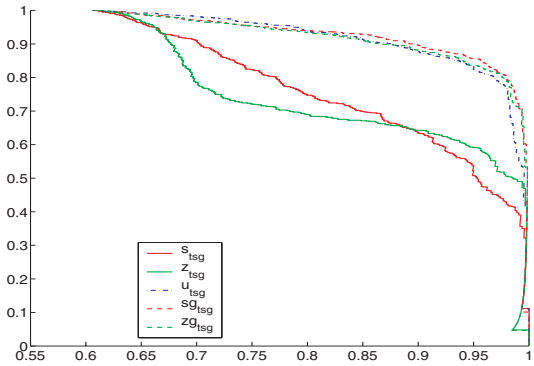


Fig. 5. Specificity–sensitivity plots for the local methods from top to bottom: ssl, sfl, tsl, tfl. Specificity is plotted on the x-axis, sensitivity on the y-axis.







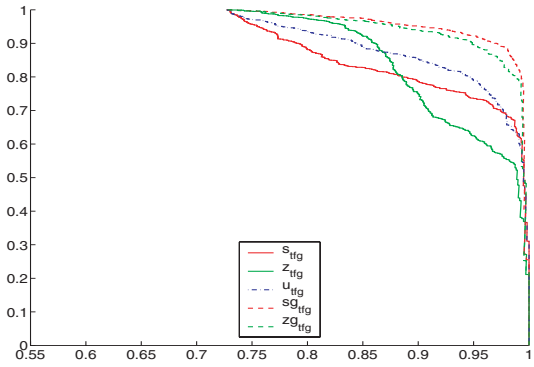


Fig. 6. Specificity–sensitivity plots for the global methods from top to bottom ssg, sfg, tsg, tfg. Specificity is plotted on the x-axis, sensitivity on the y-axis.

CONCLUSION

We evaluated the performance of different fold recognition methods for a large dataset. We find that threading with frequency profiles performs best according to our measures. For the data set analysed here, global threading performs better than local.

Further we analysed several confidence measures in order to estimate the validity of a prediction made with one of the above fold recognition methods. We find that score gaps and z-score gaps perform competitively to p-values. This is important, since high-quality confidence measures were painfully missing for global alignment and many threading and profile methods. From the confidence measures presented, we can empirically estimate the probability of a prediction being correct or incorrect. This estimate becomes essential if protein threading is used in an automated cascade of tools for protein structure prediction.

Future work includes combining several methods using appropriate confidence measures, and using frequency profiles for the structures as well which should further improve performance rates and prediction confidences.

ACKNOWLEDGEMENTS

Part of this work has been supported by the BMBF project Helmholtz Network for Bioinformatics (01SF9984/3) and DFG project Prophy (Zi 616/1).

REFERENCES

Alexandrov,N., Nussinov,R. and Zimmer,R. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Proceedings of the Pacific Symposium on Biocomputing*. pp. 53–69.

Altschul,S. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.

Altschul,S., Madden,T., Schäffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Altschul,S. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Arratia,R. and Waterman,M. (1994) A phase transition for the score in matching random sequences allowing deletions. *Bull. Math. Biol.*, **5**, 743–767.

Bailey,P., Brunak,S., Chauvin,Y., Andersen,C. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Barton,G.J. and Russell,R.B. (1993) Protein structure prediction. *Nature*, **361**, 505–506.

Benner,S.A., Cohen,M.A. and Gerloff,D. (1992) Correct structure prediction? *Nature*, **359**, 781.

Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Defay,T. and Cohen,F.E. (1995) Evaluation of current techniques for ab initio protein structure prediction. *Proteins*, **23**, 431–445.

Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K. *et al.* (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins*, Suppl **3**, 209–217.

Fischer,D., Elofsson,A. and Rychlewski,L. (2000) The 2000 Olympic Games of protein structure prediction; fully automated programs are being evaluated vis-a-vis human teams in the protein structure prediction experiment CAFASP2. *Protein Engg*, **13**, 667–670.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: Detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Henikoff,S. and Henikoff,J. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Jones,D.T. (1997) Progress in protein structure prediction. *Curr. Opin. Struct. Biol.*, **7**, 377–387.

Kallberg,Y. and Persson,B. (1999) KIND—a non-redundant protein database. *Bioinformatics*, **15**, 260–261.

Karlin,S. and Altschul,S. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Krogh,A. and Mitchison,G. (1995) Maximum entropy weighting of aligned sequences of protein or DNA. In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, California 94025: AAAI Press, pp. 215–221.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.

Mott,R. (2000) Accurate formula for p-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, **300**, 649–659.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**.

Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

von Öhsen,N. and Zimmer,R. (2001) Improving profile–profile alignment via log average scoring. In Gascuel,O. and Moret,B.M.E. (eds), *Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 2001, Proceedings*, Lecture Notes in Computer Science 2149, Springer, Berlin Heidelberg, New York, pp. 11–26.

Waterman,M. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Zien,A., Zimmer,R. and Lengauer,T. (2000) A simple iterative approach to parameter optimization. *J. Comp. Biol.*, **7**, 483–501.

APPENDIX A: FITTING DISTRIBUTIONS TO DATA

In this article we fitted Gauss and Gumbel probability distribution functions to data $\boldsymbol{x} = [x_1, \ldots, x_n]$ as follows.

Gauss distributions. are fitted by estimating μ as $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and σ as $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Gumbel distributions. $F(x) = \exp\left\{-\exp\left\{-\frac{x-a}{b}\right\}\right\}$, are fitted by estimating the parameters a and b by $\hat{b} = \frac{\sqrt{6}}{\pi}\hat{\sigma}$, and $\hat{a} = \hat{\mu} - \hat{b}\gamma$, with the Eulerian number

$$\gamma = \left\{\lim_{n\to\infty}\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n\right)\right\} \approx 0.577216.$$