

Accurate Formula for P-values of gapped local sequence and profile alignments

Richard Mott
Wellcome Trust Centre for Human Genetics
Roosevelt Drive Oxford OX3 7BN UK
Richard.Mott@well.ox.ac.uk

May 31, 2000

Abstract

A simple general approximation for the distribution of gapped local alignment scores is presented, suitable for assessing significance of comparisons between two protein sequences or a sequence and a profile. The approximation takes account of the scoring scheme (ie gap penalty and substitution matrix or profile), sequence composition and length. Use of this formula means it is unnecessary to fit an extreme-value distribution to simulations or to the results of data-bank searches. The method is based on the theoretical ideas introduced in (Mott & Tribe, 1999). Extensive simulation studies show that score-thresholds produced by the method are accurate to within $\pm 5\%$ 95% of the time. We also investigate factors which affect the accuracy of alignment statistics, and show that any method based on asymptotic theory is limited because asymptotic behaviour is not strictly achieved for many real protein sequences, due to extreme composition effects. Consequently it may not be practicable to find a general formula that is significantly more accurate until the sub-asymptotic behaviour of alignments is better understood.

Keywords: statistical significance, protein sequence, protein profile, sequence alignment

Running Title: Gapped Alignment P-values

1 Introduction

The problem of how to determine the statistical significance of sequence alignment scores has attracted much investigation, as it is of central importance to know whether an observed sequence similarity could imply a functional or evolutionary link, or is a chance event (see e.g. (Altschul & Gish, 1996; Waterman, 1995; Liu & Lawrence, 1999) for reviews of the subject). Nevertheless, we still lack a complete theoretical solution to the cases of greatest importance, that is of the optimal local gapped alignment between two sequences, (Smith & Waterman, 1981), or a sequence and a profile (Gribskov *et al.*, 1987)..

The distribution of alignment scores depends on the lengths and compositions of the sequences and on the scoring scheme. Local sequence alignments are evaluated relative to a scoring scheme in which the score for substituting the amino-acid residues a, b is $S(a, b)$, and the cost of inserting a gap of k residues is $g(k) = A + Bk$ (the affine gap penalty), or some more complicated function such as $g(k) = A + B \log k$ (Miller & Myers, 1988; Mott, 1999) . These scores may be computed using a dynamic programming algorithm (Gotoh, 1982; Mott, 1999).

The basic question to be answered is: what is the probability that a similarity score at least as great as that actually observed in a comparison between real sequences could have arisen by chance, when sampling from suitably-defined populations of random unrelated sequences? The strategies that have been used to assess significance can be classified according which factors are taken into account. This is equivalent to defining different sequence sampling populations. The simplest method is to fit a distribution to the raw scores from a search (Collins & Coulson, 1990), ignoring sequence length and composition. Next, one can take

account of sequence length but not variation in composition, by fitting a distribution to scores whilst adjusting for length (Pearson, 1998) or from a precalculated lookup-table assuming a standard composition (Altschul & Gish, 1996). Finally, one may take account of both length and composition, again either by function-fitting (Mott, 1992) or by a formula, which is the subject of this paper.

In the context of a databank search, one is interested in assessing statistical significance in order to sort the results in best-first order, and to determine which scores are likely to be significant. The order of similarities sorted by P-value will depend on the model of sequence randomness used, and clearly one should use that method which best separates the genuine similarities from the occasional high-scoring random match. Discrimination seems to correlate with how many, and how well, the confounding factors of length and composition are taken into account. (Brenner *et al.*, 1998) compared the reliability of several databank-search packages, and found that the most accurate pairwise method (SSEARCH, (Pearson, 1998)) used the Smith-Waterman algorithm with scores fitted to an extreme-value distribution, taking account of sequence length. Methods which took account of composition as well were not considered, as none were readily available at that time (function-fitting methods such as SSEARCH and FASTA implicitly take account of the query sequence's composition, but not variations between the databank sequences).

Ideally, we need a simple and easily-applied formula for the score distribution. Indeed, if no gaps are permitted in the alignments, and provided $S(a, b) < 0$ on average, then there is a theoretical asymptotic solution (Arratia *et al.*, 1988; Karlin & Altschul, 1990; Karlin & Dembo, 1992): for two random sequences of lengths m, n , the P-value of a local ungapped similarity with score at least t is

$$p = \Pr(W > t) \sim 1 - \exp(K_u m n e^{-\lambda_u t}) \quad (1)$$

where K_u, λ_u are constants depending on S and the sequences' compositions (see e.g. (Karlin & Altschul, 1990; Mott & Tribe, 1999)). The subscript u emphasizes that these quantities refer to *ungapped* alignments. λ_u is the unique positive root of the equation

$$\sum_x h(x) e^{\lambda x} = 1, \quad (2)$$

and $h(x)$ is the probability that two amino-acids drawn at random have substitution score x . The formula for K_u is more complicated but only depends on λ_u, h . This result is strictly only applicable to long sequences, and to obtain more accurate estimates of significance for the shorter sequences typically encountered, m, n are sometimes replaced by $m' = m - \ell, n' = n - \ell$, where ℓ is the expected length of a random similarity (Waterman & Vingron, 1994b; Altschul & Gish, 1996).

Of primary interest are the thresholds $t(p)$ for statistical significance at different P-values p . By solving (1) for t , we have

$$\lambda_u t(p) \sim \log K_u m n - \log(-\log(1 - p)) \quad (3)$$

$$\sim \log K_u m n - \log p, \text{ when } p \text{ is small.} \quad (4)$$

Now in a databank search in which a very large number of comparisons is made, a P-value p is not immediately useful, and should be replaced by the E-value, an estimate of the number of times such a P-value would occur by chance in the search. The E-value is equal to the P-value multiplied by N_E , the number of independent comparisons made. Because the sequence databanks are partially redundant this number is less than the actual number of sequences $N \sim 10^5 - 10^6$, so a conservative estimate of the E-value is pN . The E-value is used to obtain the threshold for significance on a data-bank basis. For example, when we choose the E-value threshold 0.01, the corresponding P-value will be somewhere in the range $10^{-8} - 10^{-6}$, depending on the size of the databank and its level of redundancy. Therefore we must be able to estimate $t(p)$ accurately when p is very small (and t large), that is, in the region where Equation (3) says that thresholds are linearly related to \log P-values. Equation (3) also indicates that when the P-value is small, it depends mainly on λ .

In future the emphasis of sequence comparisons may shift from examining of single searches to data-mining collections of all-against-all comparisons, where each of N sequences in a databank are compared against each other, or against a library of N_p profiles. In this context,

statistical significance will depend on the type of question being asked, that is, the effective database size is the number of comparisons N_c considered by the question, which can range from 1 to all the datapoints ($N_c \approx NN_p$ or $N(N-1)/2$). Consequently it will be more useful to store individual pair-wise P-values (or bit scores) and convert them to E-values on demand. For this purpose, use of a formula for computing P-values is more consistent than function-fitting, because the fitted P-value of a comparison between sequence A and profile B will depend on whether A is searched against the profiles or B against the sequences, whereas there is no difference according to a formula.

Other work (Mott, 1992; Waterman & Vingron, 1994a; Karlin & Altschul, 1993; Karlin, 1994; Altschul & Gish, 1996; Pearson, 1998; Spang & Vingron, 1998; Olsen *et al.*, 1999) strongly suggests that gapped alignment scores often have the same type of extreme-value distribution (1), but with different constants K_g, λ_g , which are usually estimated by fitting (1) to scores from simulations (Altschul & Gish, 1996; Eddy, 1998a) or databank searches (Pearson, 1998). More recently, the Greedy Extension Model (GEM), gave a new theoretical approximation for gapped scores (Mott & Tribe, 1999). In essence, the GEM replaces the optimal Smith-Waterman algorithm by a greedy, slightly sub-optimal one with similar performance on random sequences, but with simpler statistics. The distribution of GEM scores depends on the gap penalty function $g(k)$ through a parameter, α ,

$$\alpha = 2s \sum_{k>0} e^{-\lambda_u g(k)}, \quad (5)$$

which for the affine penalty $g(k) = A + Bk$ takes the simpler form

$$\alpha = 2se^{-\lambda_u(A+B)} / (1 - e^{-\lambda_u B}) \quad (6)$$

In these formulae s is closely related to K_u and is independent of the gap-penalty (Mott & Tribe, 1999). It may be shown (Stephen Altschul, personal communication) that

$$s = \sqrt{\frac{\delta K_u}{H(e^{\lambda\delta} - 1)}} \quad (7)$$

where δ is the smallest span of score values (usually 1) H is the entropy (Equation 10), and $K_u \equiv K^-$ in (Karlin & Altschul, 1990).

Under the GEM everything about the gap-penalty is captured by α , and K_g, λ_g are approximated by the factorisations

$$K_g \approx K(\alpha) = K_u \kappa(\alpha) \quad (8)$$

$$\lambda_g \approx \lambda(\alpha) = \lambda_u \theta(\alpha) \quad (9)$$

where θ, κ are functions of α alone.

Alignment behaviour varies with α as follows: When $\alpha = 0$ no gaps are allowed (corresponding to infinite gap penalties) and $K(0) = K_u, \lambda(0) = \lambda_u$. As the gap penalties decrease, α increases and gaps begin to appear, until a phase transition occurs (Waterman *et al.*, 1987). Before the transition GEM score statistics behave qualitatively like ungapped scores; P-values are calculated using $K(\alpha), \lambda(\alpha)$ substituted into (1). Post-transition alignments comprise long chains spanning the entire sequences, and follow different statistics (Waterman, 1995; Drasdo *et al.*, 1998). Simulations indicate the transition occurs in the range $0.3 < \alpha_{\text{crit}} < 0.4$. In practice, to ensure sensitivity, the scoring scheme should be such that $0 < \alpha < 0.25$ approximately, and values of $\alpha > 0.2$ should only be used in rare cases when many gaps are expected. For example, BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) and FASTA (Pearson & Lipman, 1988; Pearson, 1998), have default parameters where $\alpha \approx 0.08$ and 0.16 respectively.

For a fixed substitution matrix and pair of sequences, gapped scores are always at least as great as ungapped ones because the gapped algorithm can always return the best ungapped similarity if nothing better is found. $\theta(\alpha)$ measures the relative deflation of λ_g to λ_u , and always lies between 0 and 1. Thus $\theta(0) = 1$, θ decreases as α increases (and the gap penalty weakens), until $\theta(\alpha_{\text{crit}}) = 0$ at the phase transition.

An obvious but important consequence of the GEM is that if two gap penalty functions share the same α and substitution matrix then they will have identical random score distributions. However, they have different sensitivities (ie they need not assign a given type of similarity the same P-value; see (Mott, 1999) for an example and discussion).

For most popular scoring schemes $K(\alpha), \lambda(\alpha)$ are quite close to maximum-likelihood estimates $\hat{K}_g, \hat{\lambda}_g$, obtained by fitting the distribution (1) to Smith-Waterman scores from simulations (Mott & Tribe, 1999). Although it is feasible to use the GEM formula to compute P-values, $\lambda(\alpha)$ is consistently larger than $\hat{\lambda}_g$, and the error increases with α .

In this paper we significantly improve the accuracy of the GEM estimates, reducing bias, extending the range of α over which the estimates are reliable and incorporating sub-asymptotic length effects. Our strategy retains the factorisation in (8,9) but replaces the GEM formulae for $\theta(\alpha), \log \kappa(\alpha)$ by simple linear functions of α modified by length correction terms. We estimate the unknown linear coefficients in the formulae once only from a broad set of simulations. The result is a simple, universal approximation for P-values.

We also investigate in detail certain problems concerning the theory of ungapped alignments and the generation of random sequences, which affect the accuracy of the formula.

2 Factors which affect Score Distributions

2.1 Random Sequences and Parameter Estimation

Throughout this paper the standard model is used for sampling random protein sequences, ie the length n of the sequence is assumed fixed and the amino acid type a occurs with some known probability p_a , independent of its position in the sequence and of the neighbouring residues. Each generated sequence was shuffled as a precaution against potential serial correlation, which we observed in some random-number generators (see Discussion).

For a given scoring scheme S, g , the parameters K, λ of the score distribution may be estimated by fitting Equation (1) to the scores from a large number of comparisons between random sequences, using the method of maximum-likelihood (Mott, 1992; Waterman & Vingron, 1994b); 10000 comparisons will give an estimate of λ with a 95% confidence interval $\Delta\lambda$ of about $\pm 1.5\%$ (and $\Delta K \pm 8\%$), and 24000 comparisons about $\pm 1\%$ ($\Delta K \pm 5\%$). The number of comparisons, and hence the computer time required, may be reduced significantly by using the declumping techniques described in (Waterman & Vingron, 1994b; Olsen *et al.*, 1999), although we have not done so in this work. Most of the simulations used in this paper were performed on a farm of 25 Intel CPUs running Linux 2.2.5-15 with Mosix version 0.8.

Uncertainty $\Delta K, \Delta\lambda$ in the parameters, either from statistical fluctuations in estimates or from a misspecified model of sequence randomness, affects the accuracy Δp of the P-value as follows: For large thresholds t (ie small p), the P-value is dominated by λ , and $\Delta \log p \sim \Delta \log K - t\Delta\lambda$. To take a numerical example, for gapped alignments using standard matrices and gap penalties, ($K_g \approx 0.04, \lambda_g \approx 0.25$), the thresholds for significance at $p = 10^{-7}$ generally occur for $t(p) \approx 150$. $\Delta \log K$ only contributes a small constant effect of about ± 0.1 , independent of t , and which is negligible compared with that due to $\Delta\lambda$. Consequently, if $\Delta\lambda \sim \pm 5\%$ then $\Delta \log p \sim \pm 0.1 + 150 * 0.25 * 0.05 \approx 2.0$, and if $\Delta\lambda \sim \pm 2\%$ then $\Delta \log p \sim \pm 0.75$. In other words the P-value at the score threshold is uncertain to within a factor $e^{\Delta \log p}$, of around 2 – 10.

It should also be noted that, since most scoring schemes are integer-valued, the smallest possible change (i.e. unity) in the score will alter the P-value by a factor $e^{\pm\lambda}$, which for $\lambda \sim 0.25$ implies the P-value changes by about 30%. If the sequences were modified slightly so that just one extra match occurred in the alignment, with an additional score of about 5, then the P-value would decrease 3-4 -fold approximately. All these considerations suggest that care should be taken when interpreting the nominal significance levels given by Equation (1); they may be inaccurate by a factor of up to 10.

2.2 Onset of Asymptotic Behaviour

The theoretical results for ungapped statistics only apply when the sequences are long and have roughly similar lengths m, n (strictly, that $\frac{\log n}{\log m} \rightarrow 1$ (Arratia *et al.*, 1988)). The minimum sequence length required for the onset of asymptotic behaviour for both gapped and ungapped alignments depends on the scoring scheme and markedly on sequence composition. We have found that for random sequences with compositions close to the average, eg generated using the amino acid compositions in (Robinson & Robinson, 1991), asymptotic behaviour is achieved (i.e. the estimates of λ differ by less than 2% from theory) at quite modest lengths,

by say $m = n = 250$. However, many real sequences have quite skewed compositions, and in these cases the rate of onset is much slower; even $m = n = 1000$ may be insufficient. As the average protein length is about 330 residues, this means that we must adjust the statistics to take some account of sub-asymptotic effects.

Figure 1 illustrates how the parameter θ is affected by variation in the scoring scheme, sequence composition and sequence length. In (A), sequences were generated with standard Robinson compositions, and compared using the BLAST defaults (blosum62, 11+k). In (B) each pair of sequences was generated such that one had a composition similar to SWISSPROT entry H11L_CHICK, and the other similar to GCSH_FLATR; the scoring scheme (blosum50, 6.50+5.00k) was used. For each case, 120 sets of 10000 pairs of sequences were generated and compared. Within each set the sequence lengths were fixed, with the 120 sets covering all combinations $m \leq n$ taken from the set 75, 100, 125, 150, 175, 200, 250, 300, 400, 1000. In (B) those combinations where $n - m \geq 300$ are shown as triangles. The parameter $\hat{\theta} = \hat{\lambda}_g/\lambda_u$ was estimated by maximum likelihood for each set, and is plotted against the function $f(m, n) = \log(mn)(1/n + 1/m)$. The Figure shows that:

- θ is an increasing, approximately linear function of f .
- The asymptotic value θ_∞ , for infinitely long sequences, occurs at the intercept $f = 0$. For finite-length sequences, eg $m = n = 350$, θ is significantly larger than θ_∞ , by 6% in (A) and 22% in (B). Using θ estimated for long sequence lengths such as $m = n = 1000$ ($f = 0.0276$) for shorter lengths is also inaccurate, but the error is about half as big. Consequently, we may expect to over-estimate statistical significance if we use either asymptotic values for λ , or ones estimated from comparisons between very long sequences. This phenomenon is still present when no gaps are allowed. On average λ_u calculated using (2) are about 1.6% larger than estimates $\hat{\lambda}_u$ from simulations with sequence lengths 1000, and by extrapolation one would expect these asymptotic values to be about 2-3% too large for average-length sequences.
- The rate of onset of asymptotic behaviour depends on the slopes, which are different in (A), (B).
- In (B) those comparisons where the sequence lengths are very different (ie triangles) are significantly smaller than the others.
- For very short sequences (ie large f) $\theta > 1$, that is, $\lambda_g > \lambda_u$.

The cases (A), (B) were chosen because they represent the extremes of behaviour; most other situations should lie between them. We must therefore construct a statistical model which describes how θ (and κ) depend on $f(m, n)$, bearing in mind that the model may break down when the lengths are very different.

The choice of the function f was motivated by the edge-correction used by (Altschul & Gish, 1996), which is related to it, although the correspondence is not exact. It should be noted that functions other than $f(m, n)$ could be used equally well, e.g $(1/m + 1/n)$, $\exp(-C\sqrt{mn})$, [although $1/(m + n)$ and $\exp(-C(m + n))$ do not fit the data.]

A similar analysis of $\log \kappa$ also shows an approximately linear dependence on f . Furthermore $\theta, \log \kappa$ are almost linear functions of each other. There is no theoretical explanation for this at present.

3 Model

The first task is to model the variation in the slopes in Figure 1. It is plausible that this will depend on α (Mott & Tribe, 1999) and $1/H$, used in the length correction of ungapped alignments in (Altschul & Gish, 1996)), where

$$H = \sum_x xh(x)e^{\lambda_u x} \quad (10)$$

is the entropy of a single match. To investigate and illustrate alignment behaviour in realistic situations, throughout this paper we use a representative set of 70 pairs of protein sequences chosen at random from SWISSPROT, together with 70 combinations of substitution matrix (from the pam and blosum series) and gap penalty chosen to span the space of likely scoring

schemes. In addition 19 scoring schemes using standard protein compositions (Robinson & Robinson, 1991), including the default scoring schemes for BLASTP (blosum62, $g(k) = 11 + k$) and FASTA (blosum50, $g(k) = 10 + 2k$) were used. Figures 2,3 summarise an analysis of these 89 groups; for each group 120 sets of simulations were performed as in Figure 1, ie a total of 10680 sets of 10000 simulations. A straight-line fit $\theta = \theta_\infty + \beta f(m, n)$ was made to each group independently. Then the set of 89 slopes $\hat{\beta}$ was fitted to a linear function of $1/H, \alpha$. Figure 2 plots the observed and predicted slopes, showing good agreement. More complicated models were also tried, but did not dramatically increase the goodness of fit. The regression accounts for 92% of the variance. From the small size of the error-bars on each data point, it is clear that some of the remaining variance cannot be explained by sampling error; ie we have not captured all of the variability in our model. Note that those groups using standard (Robinson & Robinson, 1991) compositions (marked as triangles) are close to the regression line.

Turning now to the intercepts θ_∞ , one expects they should be a function $\theta(\alpha)$, which we will approximate as $c_0 + c_1\alpha$, for some constants c_0, c_1 , independent of the scoring scheme and composition; In theory $\theta_\infty(0) = 1$, so we expect $c_0 \approx 1$. Figure 3 plots the fitted values $\hat{\theta}_\infty$ against α , and shows that this model is reasonable for small α . For $\alpha > 0.2$ there is a wider dispersion round the best straight-line fit, so the model is less accurate in this region. The best-fitting line is $1.013 - 2.61\alpha$, with standard errors for the coefficients of 0.006, 0.04 respectively. Again, the error bars on the Figure indicate that some of the variation not explained by α is also not attributable to sampling error. Those groups using standard (Robinson & Robinson, 1991) compositions are close to the line of best-fit, so most of the error is due to composition effects.

Putting everything together into Equation (9), the model for λ_g is

$$\lambda_g = \lambda_u(1.013 - 2.61\alpha + f(m, n)(-0.76 + 9.34\alpha + 1.12/H)) \quad (11)$$

A similar analysis of $\log \kappa$ gives the formula for K_g of

$$K_g = K_u \exp(0.26 - 18.92\alpha + f(m, n)(-1.76 + 32.69\alpha + 192.52\alpha^2 + 3.24/H)) \quad (12)$$

Note the quadratic term in α^2 . The model for K_g is considerably less accurate than that for λ_g , but as discussed above, errors in K have relatively little effect on score thresholds.

The score thresholds $t(10^{-8})$ were computed for each of the 10680 sets, using (i) the formulae for $\theta, \log \kappa$, (ii) direct MLEs $\hat{\lambda}_g, \hat{K}_g$. Figure 4 plots the two thresholds, for those 74 groups where $\alpha < 0.2$. 95% of the predicted thresholds are within 5% of the estimated values, and 65% within $\pm 2\%$. As most of the thresholds are of the order of 100, this translates to a score uncertainty of about ± 5 , or ± 1 symbol match for most score matrices. When $\alpha > 0.2$ the model is less accurate, and its use is not recommended. However, this is rarely a handicap in practice.

3.1 Application to Real Data

The formula was tested by an analysis of the dataset PDB40D-J (Park *et al.*, 1998), comprising 935 sequences of known structure together with their SCOP structural classifications (Murzin *et al.*, 1995); 2096 of the 436,645 pairs are homologous, in the sense of having the same SCOP superfamily. This dataset is useful for benchmarking sequence comparison methods because the ‘‘true’’ relationships are known. We compared the results of our formula with SSEARCH (Pearson, 1998), which is probably the most sensitive pairwise method in current use (Brenner *et al.*, 1998). SSEARCH compares a query to a databank using Smith-Waterman, and fits an extreme-value distribution to the scores, taking account of sequence length but not variations in the sequence composition. We used the default SSEARCH scoring scheme BLOSUM50, $10 + 2k$ (i.e. 12,2 in SSEARCH terminology).

A two-step procedure was used in order to avoid computing K_g, λ_g for each comparison; Firstly an approximate P-value p_0 was obtained using the formula with fixed parameters correct for standard amino-acid composition; if $p_0 < 0.001$ then the ‘‘true’’ P-value p_1 was computed using the sequences’ compositions. In almost all cases $p_1 > p_0$, suggesting that statistics based on standard compositions over-estimate significance. P-values were converted to E-values by multiplying by 436,645.

Problems were encountered when using the observed amino-acid frequencies from very short sequences because their compositions can be skewed significantly by chance (187 sequences are shorter than 75 in PDB40D_J, and the mean length is only 171). In comparisons between two very short sequences the ungapped λ_u can drop by up to 50% from the typical value if actual frequencies are used, implying score thresholds should be doubled (and in some cases the expected match score becomes positive, so the entire extreme-value theory is inapplicable). However, this causes the false negative rate to increase dramatically. It was found that these problems were removed by adding pseudocounts for a 100-residue sequence of typical composition to the observed frequencies in each sequence; in average-length sequences the statistics are still dominated by the observed frequencies, while shorter sequences are less skewed.

In 152 comparisons where $p_0 < 0.001$ it was also found that $\alpha > 0.25$, implying the gap penalty was too weak. In these cases the penalty was increased automatically until $\alpha < 0.25$, the sequences realigned and the statistics recomputed. One advantage of using a formula to compute statistics is that one can make changes to the scoring scheme in the light of the sequences' characteristics.

Sensitivity was assessed using the method described in (Brenner *et al.*, 1998); i.e. the results were sorted by p_1 , and the fraction of homologues $F(r)$ observed at a given rate r of false positives computed (Table 1). Overall, the results from using the formula were slightly better than obtained using SSEARCH, and better than those reported for FASTA ktup=1 and WU-BLAST (Park *et al.*, 1998); they were worse than those reported for profile-based methods, as would be expected.

There was one false-positive comparison, between Tumor necrosis factor receptor Inca3, and Blood coagulation factor XA 1hcg_b_5, with E-value 0.1. Examination of the structural alignment indicated a possible homology. All other non-homologues had E-values > 1 . Moreover, the number of non-homologues observed with E-values $> N$ was always close to N (Table 1), indicating the statistical significance levels are accurate when applied to real data.

4 Statistics of sequence-profile comparisons

The emergence of curated databases of multiple alignments, such as PFAM (Bateman *et al.*, 1999), SMART (Schultz *et al.*, 1998) has meant that the methodology and practice of sequence comparison has undergone a quiet revolution. It is likely that in future novel sequence, e.g. from newly-sequenced genomes, will be routinely screened against profile libraries as well as, and perhaps eventually instead of, the protein databanks.

It is well known that the sensitivity of a search can be increased by using profiles in place of sequences (Park *et al.*, 1998). The two contexts of interest are searching (i) a protein sequence databank with a profile query (Altschul *et al.*, 1997), or (ii) a databank of profiles with a sequence query (Schaffer *et al.*, 1999). Here a profile means a sequence with a position-dependent score matrix, in which the score of aligning the amino acid a at position i in the profile is $S_i(a)$. Gaps are scored as for sequence-sequence comparisons, i.e. independent of position. We require the overall expected score to be negative. If p_a is the probability that a occurs at a given position in the sequence, then the overall probability $h(x)$ that the score between a randomly-chosen profile position and sequence residue is x is

$$h(x) = \sum_{i,a:S_i(a)=x} p_a/L \quad (13)$$

where L is the profile length.

Formally, one may substitute this distribution into the formula (2) for calculating λ_u to obtain values for ungapped statistics. However, it is not quite clear what the sampling population of the profile actually is, in particular as $L \rightarrow \infty$. We need to assume that, as the sequence and profiles lengths change, the distribution $h(x)$ remains constant. In effect, this says that the position-dependence of the profile is negligible from a statistical viewpoint, that is, it affects a relatively small number of positions. Fortunately, many profiles are characterised by having a small number of highly-conserved positions in a sea of typical sequence.

There are several systems for profile-sequence comparison in current use. HMMER (Eddy, 1998a) is based on profile hidden Markov models, where K_g, λ_g are estimated separately for

each profile from comparisons with simulated proteins of average composition. SAM (Karplus *et al.*, 1998) is also HMM-based, and uses the scores of comparisons with reversed sequences to assess significance. IMPALA (Schaffer *et al.*, 1999) uses profiles such as those generated by PSI-BLAST (Altschul *et al.*, 1997) that it rescales to make the resulting ungapped λ_u for the profile is the same as one for which results from sequence-sequence simulations using the same gap penalty are available.

In order to test whether the formulae (11,12) can be used to assess statistical significance of sequence-profile comparisons we took 1364 profiles constructed by running PSI-BLAST using PFAM-A (Bateman *et al.*, 1999) domains as seeds. For each profile, 10000 comparisons were made between a shuffled version of the profile and a randomly-generated sequence with Robinson composition, of length 350. Shuffling the sequence positions of a profile preserves the distribution h . An extreme-value distribution was fit to each set, and the results compared with those predicted by applying the formula. Figure 5 plots the predicted and observed values of score thresholds $t(10^{-8})$ and shows a generally very good agreement, with over 95% of the observations within $\pm 3\%$. In particular, the model holds even for very short domain-type profiles, (shown as triangles), where $L < 75$. There are a small number of outliers, which appear to be related to extreme profile composition, for example Topoisomerase_I (labelled T in the Figure), whose consensus sequence is 17 % K, compared with about 6% found in most sequences. The probable reason that the model fits this data-set slightly better than the training set is that the majority of these profiles were scaled such that $\lambda_u \sim 0.31, \alpha \sim 0.07$, ie well away from the danger area $\alpha > 0.2$.

5 Discussion

We have shown that gapped alignment behaviour is well-characterised by a single parameter α in conjunction with the ungapped parameters λ_u, K_u, H . The gapped parameters λ_g, K_g can be approximated as functions of α , modified by length-correction terms for shorter sequences. The accuracy of the approximation is bounded by variation in sub-asymptotic behaviour, which depends on extremes of composition not captured completely by the parameters in the model. Further theoretical work on this area is therefore worthwhile.

The formula for protein sequence P-values is accurate for practical applications using standard substitution matrices/profiles and affine gap penalties, provided $\alpha < 0.2$. The computation overhead is very slight once K_u, λ_u are available, although it is quite expensive to obtain K_u . A solution is to filter comparisons by first calculating nominal P-values assuming standard compositions, and only recompute the P-values using the correct parameters if the nominal P-value is less than some modest cutoff such as 0.001.

Real protein sequences have a more complex structure than is encompassed by the simple random sequence model that we have used here. For example, there are slight periodic dependencies in α helices. In fact, a protein sequence can be thought of statistically as a mosaic of distinct segments corresponding to the underlying components of secondary structure. In principle a Hidden Markov model (Durbin *et al.*, 1998) can be used to describe this complex structure. However, to do this it would be necessary to estimate a large number of parameters, and effectively make a prediction of secondary structure, so it is quite hard to improve on the standard random model in practice. Nevertheless, one would like to know the likely change in λ if the random sequence model were broadened out to be more realistic. It is worth noting that several random number generators were investigated during the course of this work, and the estimate $\hat{\lambda}_g$ could vary by up to 5% depending on the generator used. However, scrambling the generated sequences removed this variability, indicating the differences were due to serial correlations between sequence positions. Consequently it is plausible that more realistic models of sequence randomness would produce similar changes in λ_g . On the other hand, the results of applying the formula to the PDB40D_J data indicate that the formula works well on average.

There are two desirable properties in any system of sequence comparison; (i) specificity, ie a low rate of false positives, (ii) sensitivity, ie a low rate of false negatives. Use of the correct statistical distribution for similarity scores enables one to control the false positive rate, but does not affect the sensitivity directly, which depends ultimately on the characteristics of the gap penalty, substitution matrix or profile used. However, many profiles are constructed automatically using software such as PSI-BLAST (Altschul *et al.*, 1997), which occasionally

makes mistakes by including false positives during the construction of the profile from a multiple alignment (Park *et al.*, 1998; Eddy, 1998b), so improvements in specificity may improve the generation of profiles, and hence ultimately sensitivity as well.

It may be possible to extend the method to arbitrary gap penalties, but it will be necessary to re-estimate the length-correction coefficients in the model in case the average alignment length changes.

The statistical tests described in this paper have been implemented in a package called *ariadne*, which is available from <http://www.well.ox.ac.uk/~rmott/ariadne.html>. However, the integration of this method into third-party software is encouraged; it is straightforward to do this for any local sequence, profile or HMM alignment scheme which uses affine gap penalties independent of position and score matrices/profiles with negative expected score.

Acknowledgments

We thank Roger Tribe and Stephen Altschul, Chris Ponting and Richard Copley for helpful discussions, David Jones for providing the profiles used in the paper.

References

- Altschul, S. F. & Gish, W. (1996). Local Alignment Statistics. *Meth. Enzym.* **266**, 460–480.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
- Arratia, R. A., Morris, P., & Waterman, M. S. (1988). Stochastic scrabble: large deviations for sequences with scores. *J. Appl. Probab.* **25**, 106–119.
- Bateman, A., Birney, E., Eddy, R. D. S., Finn, R., & Sonnhammer, E. (1999). Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucl Acids Res.* **27**, 260–262.
- Brenner, S. E., Chothia, C., & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.
- Collins, J. F. & Coulson, A. F. (1990). Significance of protein sequence similarities. *Meth. Enzym.* **183**, 474.
- Drasdo, D., Hwa, T., & Lassig, M. (1998). A statistical Theory of Sequence Alignment with Gaps. In: *Sixth International Conference on Intelligent Systems for Molecular Biology* pp. 52–58, Menlo Park, CA: AAAI Press.
- Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis. Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.
- Eddy, S. R. (1998a). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Eddy, S. R. (1998b). Multiple-alignment and sequence searches. *Trends Guide to Bioinformatics*, pp. 15–18.
- Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol*, **162**, 705–708.
- Gribskov, M., McLachlan, A., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**(13), 4355–4358.
- Karlin, S. (1994). Statistical studies of biomolecular sequences: score-based methods. *Philos Trans R Soc Lond B Biol Sci*, **344**, 391–402.
- Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.
- Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.* **90**, 5873–5877.

- Karlin, S. & Dembo, A. (1992). Limit distributions of maximal segmental score amongst Markov-dependent partial sums. *Adv. Appl. Prob.* **24**, 113–140.
- Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologues. *Bioinformatics*, **14**, 846–856.
- Liu, J. & Lawrence, C. (1999). Bayesian inference on biopolymer models. *Bioinformatics*, **15**(1), 38–52.
- Miller, W. & Myers, E. W. (1988). Sequence Comparisons with Concave Weighting Functions. *Bulletin of Mathematical Biology*, **50** (2), 97–120.
- Mott, R. & Tribe, R. (1999). Approximate Statistics of Gapped Alignments. *J. Comp. Biol.* **6**, 91–112.
- Mott, R. F. (1992). Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.
- Mott, R. F. (1999). Local Sequence Alignments with Monotonic Gap Penalties. *Bioinformatics*, **15**(6), 455–462.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
- Olsen, R., Bundschuh, R., & Hwa, T. (1999). Rapid Assessment of Extremal Statistics for Gapped Local Alignment. In: *Seventh International Conference on Intelligent Systems for Molecular Biology* pp. 303–309, Menlo Park, CA: AAAI Press.
- Park, J., Karplus, K., Barrett, C., Haughey, R., Haussler, D., Hubbard, T., & Chothia, C. (1998). Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods. *J. Mol. Biol.* **284**, 1201–1210.
- Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.
- Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**(2), 2444–2448.
- Robinson, A. & Robinson, L. (1991). Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. USA*, **88**(20), 8880–8884.
- Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., & Altschul, S. F. (1999). Software to Match a Protein Sequence Against a Collection of PSI-BLAST-Constructed Position-Specific Score Matrices. *Bioinformatics*, **15**, 1000–1011.
- Schultz, J., Milpetz, F., Bork, P., & Ponting, C. (1998). SMART, a simple modular architecture research tool: Identification of signalling domains. *Proc. Natl. Acad. Sci. USA*, **95**, 5857–5864.
- Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
- Spang, R. & Vingron, M. (1998). Statistics of large-scale sequence searching. *Bioinformatics*, **14**, 279–284.
- Waterman, M. S. (1995). *Introduction to Computational Biology. Maps sequences and genomes*. London: Chapman and Hall.
- Waterman, M. S., Gordon, L., & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. USA*, **84**, 1239–1243.
- Waterman, M. S. & Vingron, M. (1994a). Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. USA*, **91** (11), 4625–4628.
- Waterman, M. S. & Vingron, M. (1994b). Sequence Comparison Significance and Poisson Approximation. *Stat. Sci.* **9**, 367–381.

Figures

Figure 1

Maximum-likelihood estimates $\hat{\theta}$ vs $f(m, n)$ for two instances A, B, of scoring scheme and composition (see text), each representing 120 sets of 10000 simulations, over a range of sequence lengths m, n . The error bars are the standard errors of the estimates. The triangles represent those sets in (B) where $|m - n| > 300$.

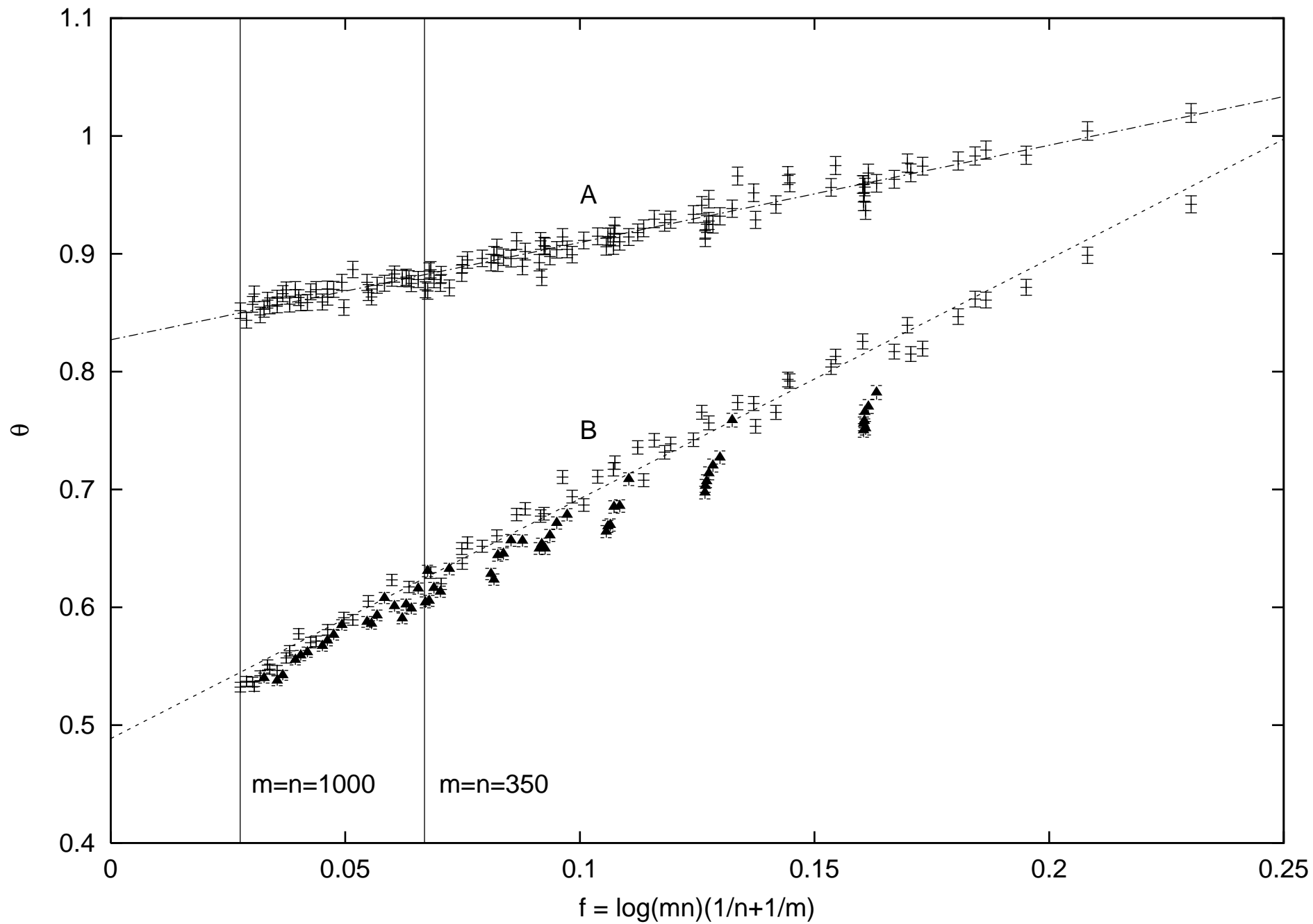


Figure 2

Fitted and observed slopes β , measuring the rate of onset of asymptotic behaviour for 89 different scoring schemes/compositions. Each data point is a pair $[\hat{\beta}, -0.76 + 9.34\alpha + 1.12/H]$ representing the observed and fitted slopes from 120 simulation sets. The points labelled A, B refer to the data in Figure 1. The error bars represent the standard errors of the $\hat{\beta}$. The triangles indicate those groups with standard sequence composition taken from (Robinson & Robinson, 1991).

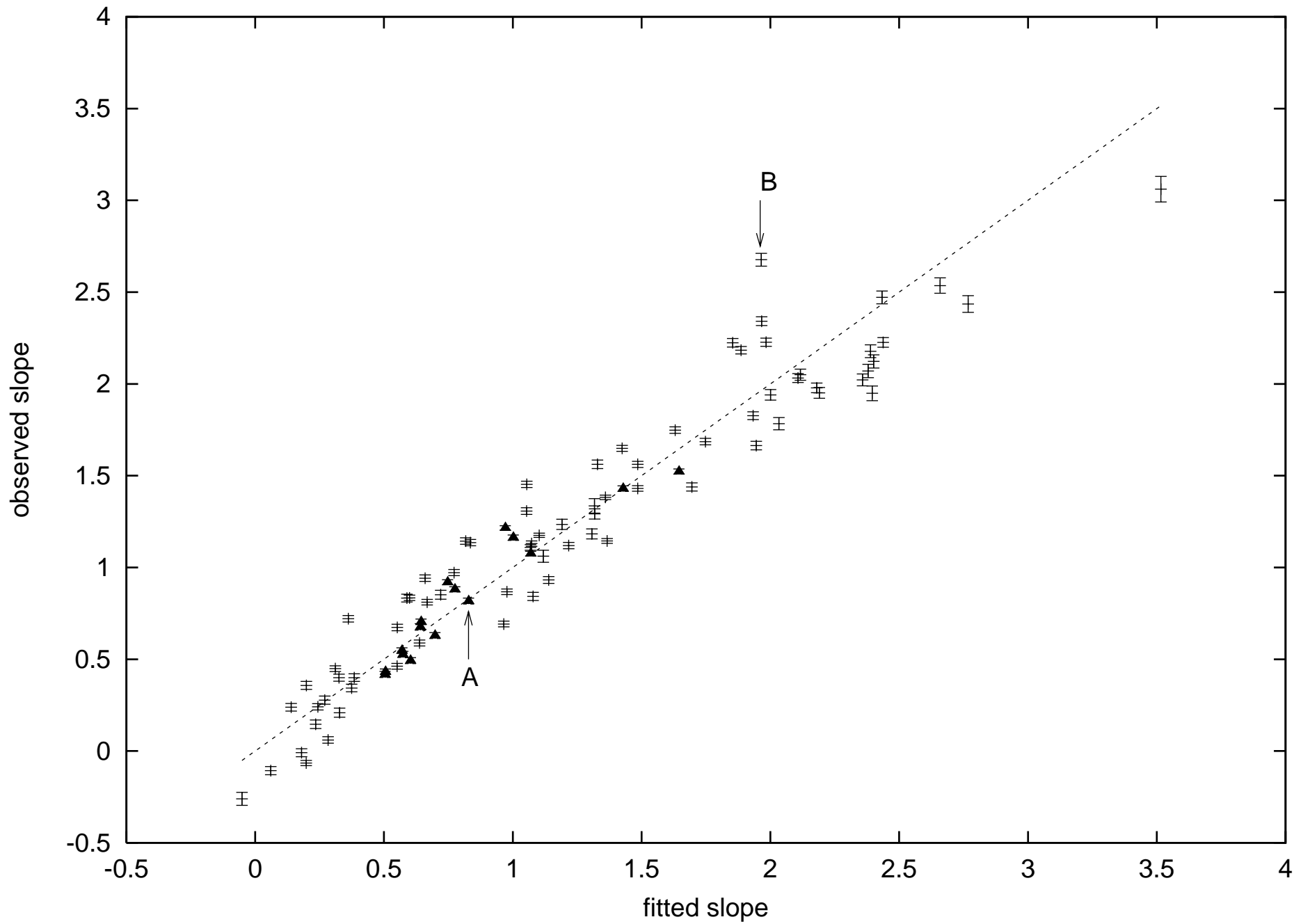


Figure 3

The extrapolated asymptotic values (intercepts) θ_∞ vs α for the 89 groups. The straight line is the least-squares regression line $1.013 - 2.61\alpha$. The points labelled A, B refer to the data in Figure 1. The error bars represent the standard errors of $\hat{\theta}_\infty$. The triangles indicate those groups with standard sequence composition taken from (Robinson & Robinson, 1991).

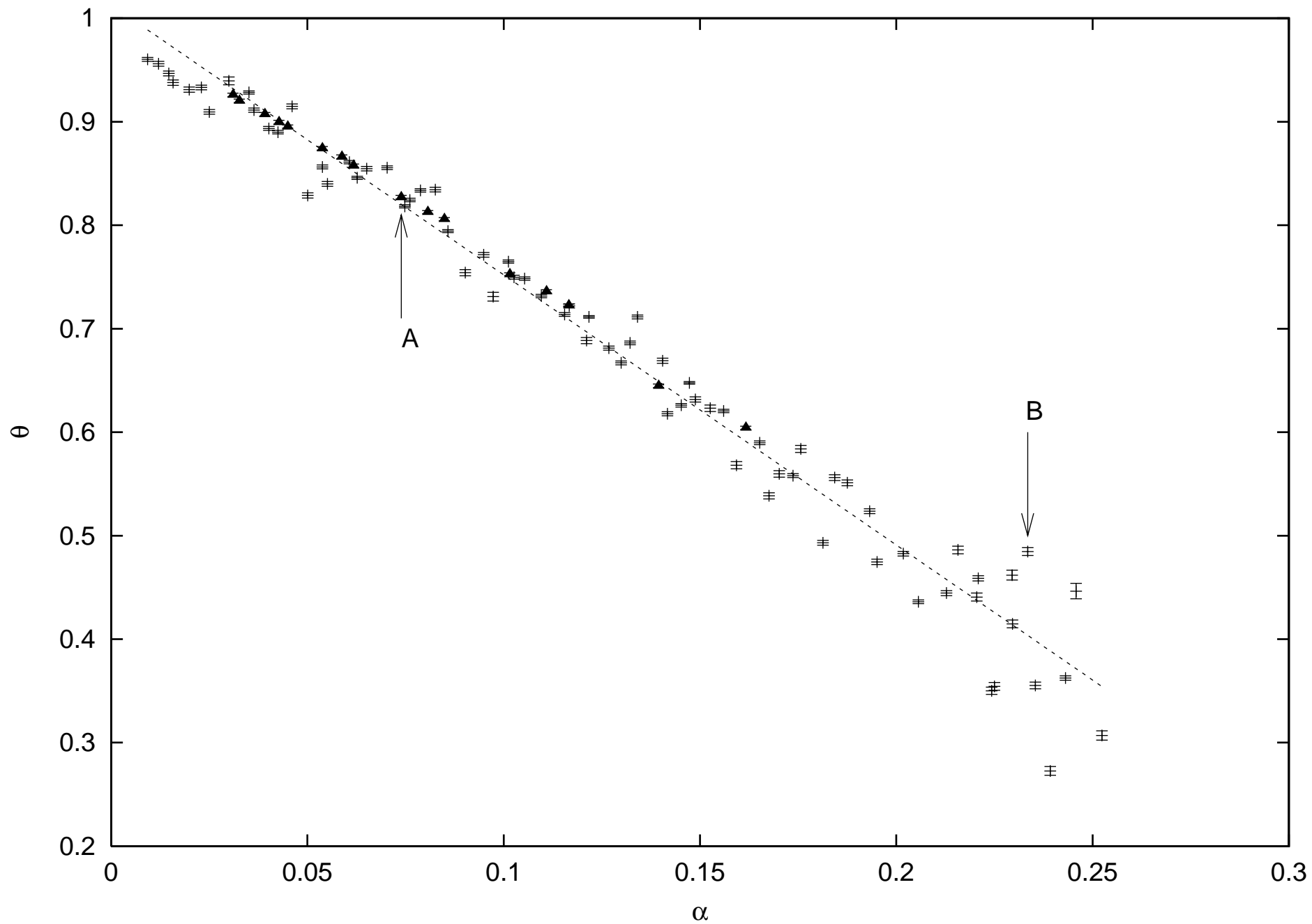


Figure 4

The score thresholds $t(10^{-8})$ for statistical significance at P-value 10^{-8} , for the 74 groups (8880 data sets) for which $\alpha < 0.2$. Y-axis: using $\hat{K}_g, \hat{\lambda}_g$ estimated directly from the 10000 comparisons in each data set; X-axis: using the formulae (12,11) for K_g, λ_g . The three straight lines $y = 1.0x, 0.95x, 1.05x$ show the line of perfect agreement and for $\pm 5\%$ error.

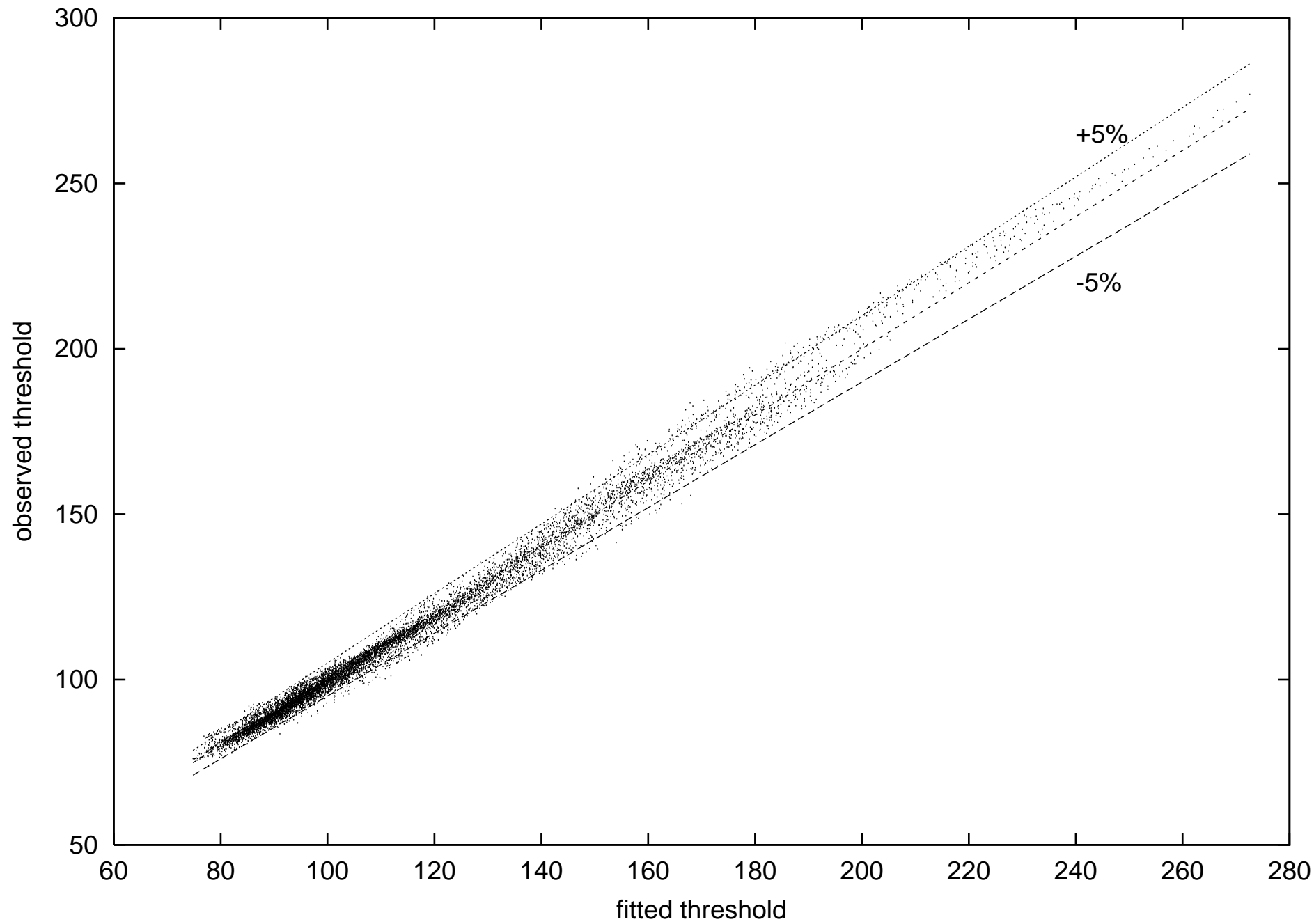
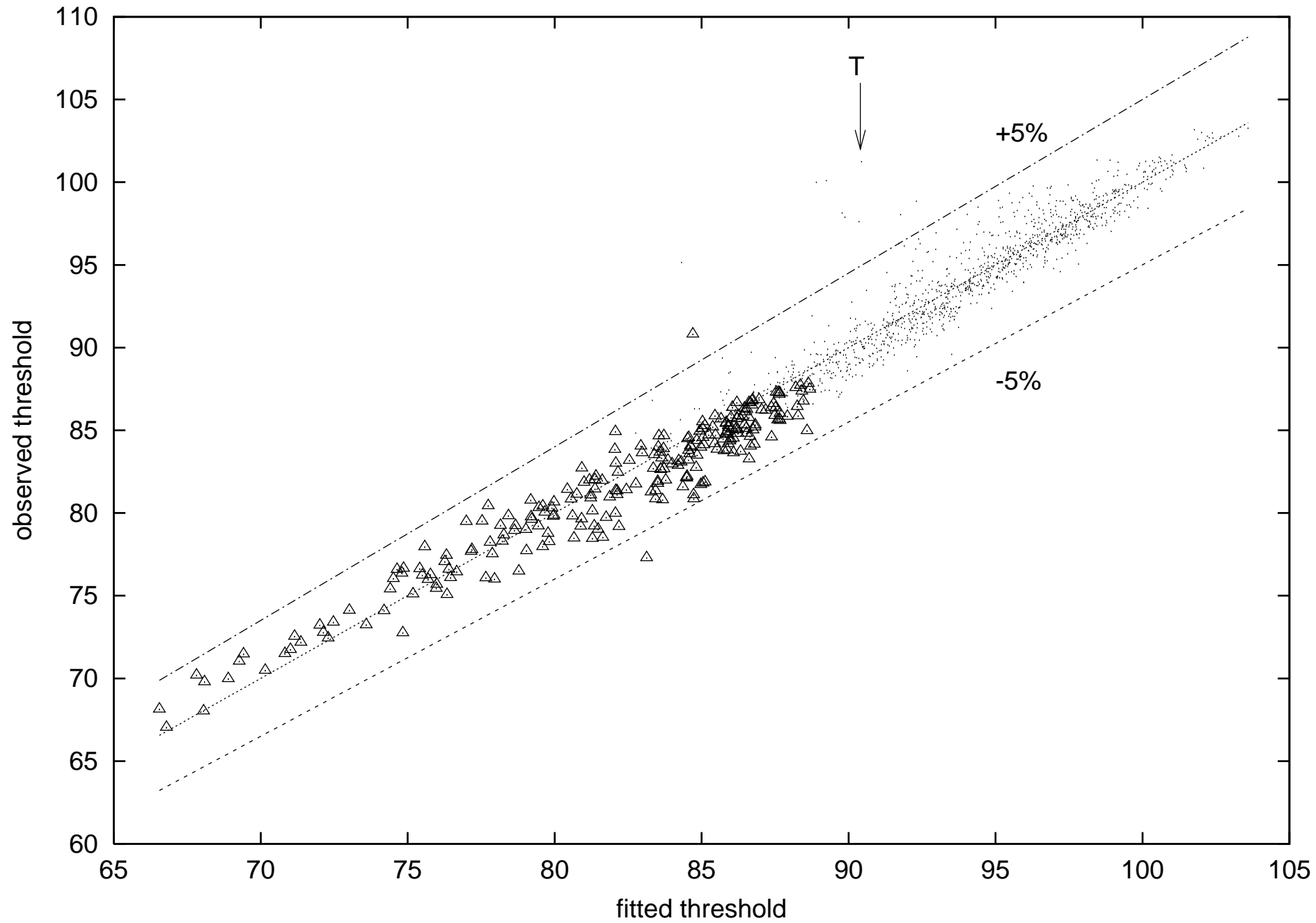


Figure 5

The score thresholds $t(10^{-8})$ for statistical significance at P-value 10^{-8} , comparisons between 1364 PSI-BLAST profiles and random sequences of length 350. Y-axis: using $\hat{K}_g, \hat{\lambda}_g$ estimated directly from the 10000 comparisons in each data set; X-axis: using the formulae (12,11) for K_g, λ_g . The three straight lines $y = 1.0x, 0.95x, 1.05x$ show the line of perfect agreement and for $\pm 5\%$ error. The outlier labelled (T) corresponds to the profile Topoisomerase.I. Triangles indicate profiles shorter than 75 positions.



Tables

Table 1

The number of homologous matches made between PDB40D_J sequences at different rates of false positives using SSEARCH and the Formula to determine statistical significance. The Predicted Errors are the largest reported E-values among the false positives found at the corresponding Observed Error rate; these should be close to the Observed Errors if the estimates of statistical significance are accurate.

False Positive Rate	Observed Errors	SSEARCH		Formula	
		Homologous Pairs	Predicted Errors	Homologous Pairs	Predicted Errors
1/100000	4	340	6	362	6
1/50000	9	371	16	379	11
1/10000	43	425	70	426	50
1/5000	87	447	93	437	84
1/1000	437	520	467	512	408