BIOINFORMATICS

Vol. 18 no. 6 2002
Pages 864–872

# Hybrid alignment: high-performance with universal statistics

Yi-Kuo Yu[1], Ralf Bundschuh[2,*] and Terence Hwa[2]

[1]Department of Physics, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431-0991, USA and [2]Department of Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0319, USA

Received on April 20, 2001; revised on January 8, 2002; accepted on January 22, 2002

## ABSTRACT

The score statistics of a recently introduced 'hybrid alignment' algorithm is studied in detail numerically. An extensive survey across the 2216 models of protein domains contained in the Pfam v5.4 database (Bateman *et al.*, *Nucleic Acids Res.*, **28**, 263–266, 2000) verifies the theoretical predictions: For the position-specific scoring functions used in the Pfam models, the score statistics of hybrid alignment obey the Gumbel distribution, with the key Gumbel parameter $\lambda$ taking on the asymptotic value 1 universally for all models. Thus, the use of hybrid alignment eliminates the time-consuming computer simulations normally needed to assign $p$-values to alignment scores, freeing the users to experiment with different scoring parameters and functions. The performance of the hybrid algorithm in detecting sequence homology is also studied. For protein sequences from the SCOP database (Murzin *et al.*, *J. Mol. Biol.*, **247**, 536–540, 1995) using uniform scoring functions, the performance is found to be comparable to the best of the existing methods. Preliminary results using the PfamA database suggest that the hybrid algorithm achieves similar performance as existing methods for position-specific scoring systems as well. Hybrid alignment is thereby established as a high performance alignment algorithm with well-characterized, universal statistics.

**Contact:** yyu@fau.edu

## 1 INTRODUCTION

Due to the rapid growth of the DNA and protein databases, computer-assisted sequence alignment tools have become an integral part of modern molecular biology. Automated tools such as BLAST (Altschul *et al.*, 1990, 1997) and FASTA (Pearson, 1988) align each query sequence with those in the database and quantify their mutual similarity by an alignment score and a $p$-value. The latter is the probability of obtaining a score of the observed value or higher 'by chance', and is generally a more meaningful measure of the significance of an alignment than the alignment score itself. The usefulness of the $p$-value is of course dependent upon the choice of null models. One common choice is simply a random amino acid sequence drawn according to some fixed background frequencies $p(a)$ for amino acid $a$. This is for example the approach adopted by BLAST, as well as by more specific tools such as HMMer (Eddy *et al.*, 1995; Eddy, 1995).

Theoretical understanding of the score statistics for this null model is well developed for local alignment algorithms that forbid insertions and deletions (or indels), the so-called 'gapless' alignments. For such alignments, the null statistics of the alignment score $\mathsf{S}$ is shown (Karlin and Altschul, 1990) to follow the Gumbel distribution (Gumbel, 1958)

$$\Pr(\mathsf{S} < x) = \exp(-KMNe^{-\lambda x}) \qquad (1)$$

in the limit that the lengths $M$ and $N$ of the aligned sequences are large. Moreover, there exist explicit formulae to compute the two Gumbel parameters $\lambda$ and $K$, for a large class of substitution matrices each containing over 200 parameters, and for arbitrary amino acid background frequencies $p(a)$.

Unfortunately, gapless alignment is not sensitive enough for detecting distant homologs where insertions and deletions with respect to the query sequence are anticipated (Brenner *et al.*, 1998). For alignment algorithms which allow gaps, e.g. the Smith–Waterman algorithm (Smith and Waterman, 1981) and BLAST (Altschul *et al.*, 1997), the null statistics are known empirically to obey Gumbel statistics as well (Smith *et al.*, 1985; Collins *et al.*, 1988; Mott, 1992; Waterman and Vingron, 1994a,b; Altschul and Gish, 1996; Olsen *et al.*, 1999). However, the dependence of the two Gumbel parameters $\lambda$ and $K$ on the hundreds of scoring parameters (including gap costs) are not known. There are some recent theoretical developments in understanding gapped alignment statistics, for special scoring functions (Bundschuh, 2000), or when only a few gaps are allowed (Siegmund and Yakir, 2000); there also exist heuristics applicable to the regime of large

[*] Present address: Department of Physics, The Ohio State University, 174 West 18th Avenue, Columbus, OH 43210-1106, USA

gap costs (Mott and Tribe, 1999; Mott, 2000). However, as these developments are still not sufficiently general for real applications, the null statistics is usually determined by large simulations. Because the simulations are costly (BLAST determines the null statistics by aligning 24 000 pairs of random sequences of length 1000 for each set of scoring parameters, one is limited to doing alignment with a small set of pre-selected scoring parameters and also with fixed amino acid frequencies $p(a)$, regardless of the actual composition of the query sequence.

It has been realized for some time now that more sensitivity can be gained if one uses position-specific scoring functions based on other information, e.g. multiple alignments, or the 3D protein structure. This is the strategy implemented in the existing high-performance tools such as PSI-BLAST (Altschul *et al.*, 1997), HMMer (Eddy *et al.*, 1995; Eddy, 1995) and SAM (Karplus *et al.*, 1998). Unfortunately, the null statistics is even less understood for alignment with position-dependent scoring functions. For this reason, PSI-BLAST limits itself to uniform gap cost and only allows the substitution scores to be position-specific[†], thereby potentially compromising the performance. HMMer does allow for position-specific gaps and evaluates the null statistics by simulating each of the more than 2000 protein domain models it uses, followed by fits to Gumbel distributions. However, we will show below that the null statistics of HMMer is actually not of the Gumbel form, hence making the reported statistics inaccurate. SAM adopts the probabilistic hidden markov models which can be viewed as probabilistic local alignment (Bucher and Hofmann, 1996) with position-specific scoring functions. Yet, the null statistics for probabilistic local alignment is not well understood either. Empirically, the score distribution is found to exhibit a form that is not described by any simple known function over the regime examined (Yu and Hwa, 2001).

To overcome the null statistics problem, (Yu and Hwa, 2001) recently proposed to modify the alignment algorithm in order to make the statistics tractable. They introduced a new algorithm which is a hybrid of Smith–Waterman and probabilistic local alignment. It was predicted that the null statistics of the hybrid alignment obeys the Gumbel form, with the important Gumbel parameter $\lambda$ taking on a fixed value of 1 for a wide range of scoring functions/parameters, including the position-specific gap costs, and for arbitrary background amino acid frequencies $p(a)$. The prediction was verified numerically for certain uniform (i.e. position-independent) scoring functions with affine gap cost. Here, we extend the study to position-specific scoring functions. Using the 2216 models of protein domains in the Pfam v5.4 database (Bateman *et al.*, 2000), we show that the null statistics of hybrid alignment still obeys Gumbel statistics with asymptotic $\lambda = 1$ as predicted. We further evaluate the performance (i.e. sensitivity) of the hybrid alignment, using protein sequences from the SCOP database (Murzin *et al.*, 1995; Brenner *et al.*, 1996). We find the performance of the hybrid alignment in the SCOP database to be similar to the best of the existing algorithms that use uniform scoring functions. Preliminary results on a sequence model from the Pfam database suggest that the performance of hybrid alignment also compares favorably to existing algorithms for position-specific scoring systems. We thereby establish that the hybrid alignment is a high-performance alignment algorithm with well-characterized null statistics. Since hybrid alignment drastically reduces the time needed to assess the statistical significance for arbitrary user-specified scoring functions[‡], it allows the users to experiment with different scoring functions (in particular gap functions) and evaluate their performance in real time. It also allows adaptation of the score statistics to account for the amino acid composition of individual query sequences. A preliminary BLAST implemenation of the hybrid algorithm is available upon request to the authors.

In what follows, we will first describe the null statistics study in which we apply hybrid alignment to the Pfam models. Then, we present the performance evaluation. We will give the details of the hybrid algorithm for both the uniform and position-specific scoring functions in Appendix A.

## 2 SCORE STATISTICS FOR POSITION-SPECIFIC SCORING FUNCTIONS

In this section, we examine numerically the predicted universality of the null statistics of hybrid alignment for biologically relevant position-specific scoring functions. Specifically, we use the 2216 models of protein domains contained in the Pfam database v5.4 (Bateman *et al.*, 2000). Each of these models is constructed carefully, starting from a multiple alignment of some manually chosen core members of each protein family. The Pfam models are cast in terms of hidden Markov models, which specify the probability of substitution, insertion, and deletion at each 'node' of the model. Taking the logarithm of these probabilities converts them into insertion, deletion, and substitution scores. These scores strongly depend on the position within the model and reflect the functional context of the corresponding amino acid. For example, nodes with very negative substitution scores correspond to the highly conserved regions, i.e. they require specific amino acids, and the protein exterior is characterized by nodes with low insertion/deletion costs.

[†] Note that the effect of the position-specific substitution score in *gapless* alignment is well-understood from the theory of Karlin and Altschul (1990).

[‡] The other Gumbel parameter $K$ can be efficiently determined once $\lambda$ is known; see Appendix B.

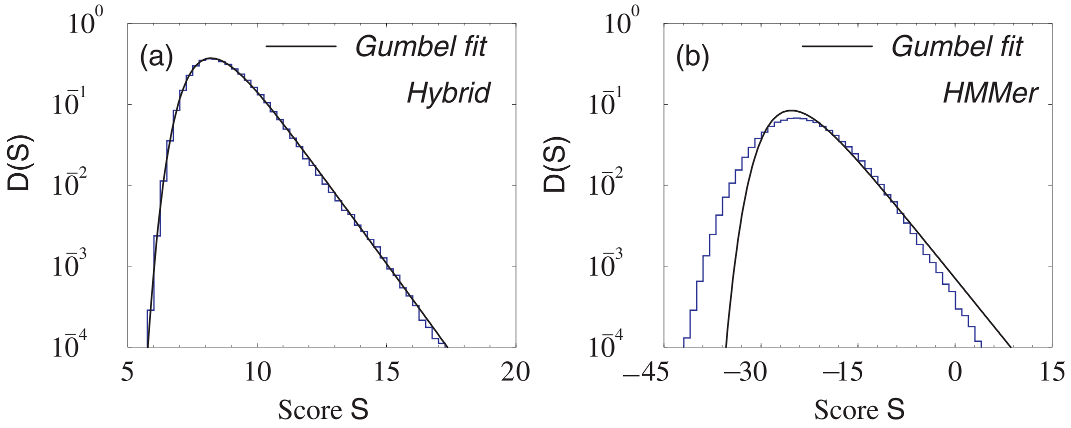

**Fig. 1.** The score distributions (staircases) of (a) hybrid alignment and (b) the sum-of-all-paths algorithm of HMMer, and their respective best Gumbel fits (solid lines). The position-specific scoring function used is the EGF protein family profile according to Pfam. The score distributions are obtained via aligning 300 000 random sequences of length 300 against the EGF profile which has a length of 45 nodes.

We first want to determine the form of the score distribution very precisely. Since this is a very challenging task numerically, we choose below the EGF protein family for a detailed study. We generate 300 000 random amino acid sequences each of length 300, according to the amino acid background frequencies specified in the EGF model. Each of the random sequences is then aligned with the EGF model using the position-specific hybrid algorithm as specified in Appendix A [Equation (17) augmented by Equations (3)–(5), and Table 2]. The resulting alignment score S is recorded for each pair. The score distribution $D(\mathsf{S})$ is obtained by normalizing the histogram of the 300 000 alignment scores, and is shown as the staircase in Figure 1(a). The empirical distribution is then fitted to the Gumbel form Equation (1) using the maximum likelihood method (Eddy, 1997). The result, shown as the smooth line in Figure 1(a), clearly describes the empirical distribution very well. The Gumbel parameters extracted from the fit are $\lambda = 1.0085 \pm 0.005$ and $K = 0.294$. Thus, our prediction that the score distribution is of the Gumbel form with $\lambda = 1$ (up to small corrections for the sequence length to be discussed in Appendix B) is verified for this particular model.

Incidentally, we note that a similar study using the sum-of-all-paths version of the algorithm of HMMer (Eddy *et al.*, 1995; Eddy, 1995) produces a score distribution (staircase in Figure 1(b)) which deviates significantly from the best Gumbel fit (thick solid line). Deviations of similar magnitudes have also been observed for a number of other models examined, including the rrm, ig, and some 7tm families. The best Gumbel fit overestimates the probability tails of some of these models and underestimates others. At present, a systematic understanding of the HMMer score statistics is lacking.

We ideally want to perform such detailed characterizations for all of the more than 2000 models in the Pfam database. However, this is not feasible within a reasonable amount of computer time. Thus, we take it for granted from the example of the EGF family shown above and a number of other families (not shown) that the shape of the distribution is always of the Gumbel form. What we can do within a reasonable amount of time is determining the Gumbel parameter $\lambda$ for all the Pfam models. To this end, we align each Pfam model to only 5000 random sequences (of length 325) using the hybrid algorithm. Again, the score distribution obtained is fitted to the Gumbel form in the same way, and a $\lambda$ value is extracted for each Pfam model. The smaller number of random sequences used reduces the accuracy on the $\lambda$s to about $\pm 1\%$ but it allows us to survey the entire Pfam database. A histogram of these $\lambda$ values obtained each of the 2216 Pfam v5.4 database is shown in Figure 2. Clearly, the distribution of $\lambda$ values is peaked at $\lambda = 1$ as predicted, with a mean of 1.014 and a standard deviation of 0.036. In comparison, a similar study using HMMer (with forced Gumbel fits) finds $\lambda$ to lie in a wide range, between 0.02 and 0.99, with a mean value of 0.33 and a standard deviation of 0.23.

The small deviation from $\lambda = 1$ shown in Figure 2 has several possible sources: One is simply due to the accuracy of the fit given the small sample size of 5000 random sequences. The other is a systematic shift due to the finite lengths of the sequences and the models. (Recall that the Gumbel distribution strictly holds only in the limit of very long sequences.) The systematic sequence length dependence can also be understood (Yu and Hwa, 2001). Correcting for the sequence and model length is expected to produce an even narrower distribution of the $\lambda$s and will be discussed elsewhere.

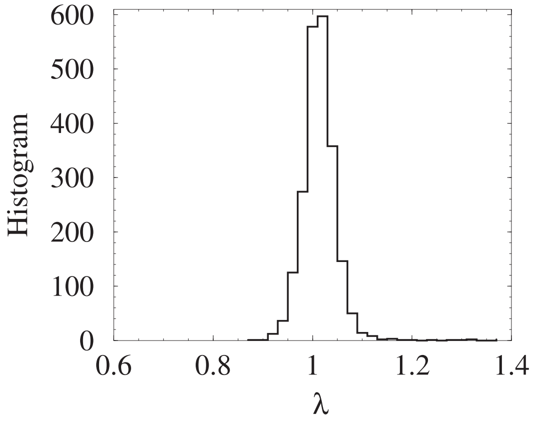

**Fig. 2.** Histogram of the Gumbel parameter λ for the hybrid alignment of random sequences to the 2216 Pfam models. The λ's are strongly peaked around 1 as predicted theoretically with a mean of 1.014 and a standard deviation of 0.036.

But even with this deviation of several percent, the advantage of hybrid alignment becomes obvious: In order to assign $p$-values to alignment scores computed by programs such as HMMer, the moderately expensive (time consuming) simulations presented above are necessary for every new or updated family. On the contrary, with hybrid alignment we can simply use a Gumbel distribution with $\lambda = 1$, without the need of much simulation. (The other Gumbel parameter $K$ can be easily determined once $\lambda$ is known; see Appendix B.) This advantage becomes especially useful when one is trying out parameters in the model-building phase and will be elaborated on elsewhere.

## 3 PERFORMANCE EVALUATION

Having a well-characterized null statistics saves the computation time otherwise necessary to obtain the statistics. However, it does not in itself make an algorithm useful. For example, the statistics of gapless alignment is well known, yet for better performance, e.g. to detect weakly related sequences, it is necessary to include gaps. What is desired is a high-performance alignment algorithm which has well-characterized statistics. Evaluating the performance of an algorithm properly is not a simple matter, because the performance does not depend only on the algorithm used but also on the appropriateness of the scoring functions, and choices of the test set and the 'gold standard'. In this section, we evaluate the performance of the hybrid algorithm, first using a uniform scoring function and then one example of a position-specific scoring function. The 'gold standard' used is the SCOP classification (Murzin *et al.*, 1995; Brenner *et al.*, 1996) in the first case, and the Pfam classification (Bateman *et al.*, 2000) in the second case. In each case, we compare the performance of the hybrid method to the corresponding 'standard' existing method using the same scoring functions. We will find that the performance of the hybrid method is similar (i.e. within statistical error) to that of the standard method in both cases. Thus, the benefit of not having to perform long simulations in order to characterize the statistics is provided by hybrid alignment at no cost in performance compared to the standard methods.

We first assess the sensitivity of the hybrid algorithm in detecting homology among protein sequences contained in the SCOP database (Murzin *et al.*, 1995; Brenner *et al.*, 1996) for which detailed structural information (and hence true similarity) is known. We largely follow the study of Brenner *et al.* (1998) who did a similar performance evaluation for a number of existing algorithms, e.g. WU-BLAST, FASTA, and gapless BLAST. Specifically, we use two subsets of the SCOP sequence domains that Brenner *et al.* (1998) selected after removing redundancy: The PDB90D–B sequences which are no more than 90% identical to each other, and the PDB40D-B sequences which are no more than 40% identical to each other. These subsets of SCOP contain 2079 and 1323 protein sequence domains, respectively. For each of the sequence sets, we calculate the alignment score between every sequence pair in the set using the hybrid algorithm. The $p$-value of the alignment score is then computed using the statistical theory described by Yu and Hwa (2001).

To evaluate the performance, we hypothesize that the two sequences $i$ and $j$ are similar if the $p$-value $p(i, j)$ of their alignment score is below some threshold $p_0$. This is to be compared to the 'superfamily' classification of the SCOP database which we take as the 'gold standard'. The superfamily classification separates the sequence pairs into $N_p$ pairs which are truly similar and $N_n$ pairs which are not similar. It also separates the pairs deemed similar by the alignment algorithm (i.e. pairs with $p(i, j) < p_0$) into two classes: Those pairs which are also similar according to the gold standard are called 'true positives', while the rest are called 'false positives'. Clearly, the numbers $\mathrm{tp}(p_0)$ and $\mathrm{fp}(p_0)$ of true and false positives respectively depend on the choice of the threshold $p_0$. Following Gribskov and Robinson (1996), we will characterize the algorithm by its coverage rate $c(p_0) \equiv \mathrm{tp}(p_0)/N_p$ and its false positive rate $f(p_0) \equiv \mathrm{fp}(p_0)/N_n$. The plot of $c(p_0)$ against $f(p_0)$ is known as the Receiver Operating Characteristic (ROC) curve. The overall sensitivity is defined as the area under the ROC curve, i.e. $\int c(f)\mathrm{d}f$.

In Figure 3(a) and (b), we show the ROC curves (solid lines) as obtained with the PDB40 and PDB90 sequences respectively, using the hybrid alignment algorithm (Equations (2)–(5) of Appendix A) with the BLOSUM-62 substitution matrix as given by Henikoff and Henikoff (1992)

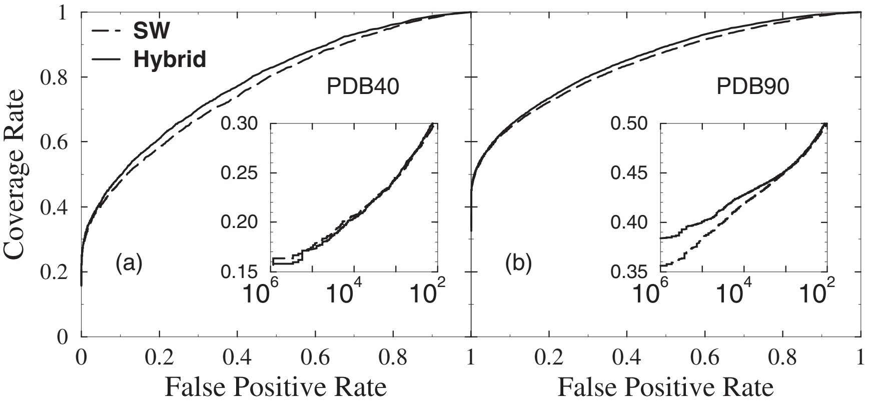

**Fig. 3.** Plots of ROC curves for both the hybrid algorithm (solid line) and the Smith–Waterman algorithm (dashed line) for (a) the PDB40D–B database and (b) the PDB90D–B database. We find that the hybrid algorithm in general performs comparable to the Smith–Waterman algorithm. The insets magnify the regions very close to zero false-positive rate; they correspond to the regions studied in detail by Brenner *et al.* (1998).

and the $11 + k$ affine gap cost. The sensitivities of the hybrid algorithm on the two databases are 0.791 for PDB-40 and 0.855 for PDB-90. A comparison of these ROC curves to the corresponding results in the study by Brenner *et al.* (1998) reveals that the performance of hybrid alignment is comparable to the best of the algorithms tested there.

To do a more quantitative comparison, we repeat the above process using the Smith–Waterman algorithm[§], which is generally recognized as the best among the existing algorithms. We use the same BLOSUM-62 substitution matrix and $11 + k$ affine gap function. Here, the conversion from alignment scores to the $p$-values requires the knowledge of the score distribution function which can only be obtained from large simulations: For each of the five different lengths ($N = 75$, 150, 300, 600, and 900) 50 000 random sequence pairs were generated and aligned in order to obtain reliable score distributions which enable accurate Gumbel fits including length corrections.

The resulting ROC curves are plotted as the dashed lines in Figure 3(a) and (b). It is clear that the performance of the hybrid and Smith–Waterman algorithms are comparable. Small differences in the ROC curves and in the sensitivity measure (0.767 for PDB-40 and 0.844 for PDB-90) are not deemed significant, as arbitrary removal of a subset of the sequences results in changes that are of the same order as the observed differences. The same qualitative result, that the performance of hybrid and Smith–Waterman alignment are comparable, is obtained for a number of other (uniform) scoring functions we examined, although the absolute performance may differ. For example, the use of PAM substitution matrices (Dayhoff *et al.*, 1978) led to worse performance by both algorithms.

Finally, we repeat the above procedure to evaluate the performance of the hybrid alignment with position-specific scoring functions. Due to the computational complexity of the task, a thorough survey of the performance of hybrid alignment for different position-specific scoring systems has not yet been performed. Instead, we present as a preliminary result the detailed study of one specific scoring function, namely the Pfam model for the immunoglobulin (ig) domain. We align the model to all of the sequences (over a quarter of a million) in the Pfam A database using the position-specific version of the hybrid algorithm as done in Section 2. The $p$-value of each alignment is obtained using the Gumbel distribution with $\lambda = 1$ as verified in Section 2, with the other (less important) Gumbel parameter $K$ obtained via methods described in Appendix B. Here, we use the Pfam classification as the 'gold standard', and compute the coverage and false positive rates by comparing the putative homologs based on the $p$-value cutoff to the 5771 sequences belonging to the ig family according to Pfam v5.4. From the resulting ROC curve, we find a sensitivity measure of 0.992 for the hybrid method.

Such a high sensitivity number should not be surprising, since the ig model was built (i.e. its parameters were fine tuned) precisely to discriminate the set of sequences belonging to the ig family. We expect a similar test using the HMMer program to yield a sensitivity of 1.0, since HMMer was the very program used to build the Pfam models. Nevertheless, the high sensitivity number obtained above for the hybrid method further supports the notion that the performance of the hybrid algorithm is comparable to the best of the existing algorithms. In fact,

[§] The exact algorithm used corresponds to Equation (4) of Appendix A, together with the Viterbi version of Equations (2) and (5) with $\mu' = 1$ and $\eta = 1$.

**Table 1.** Sensitivity of hybrid alignment: The PDB40 and PDB90 sequences are tested using the BLOSUM62 substitution matrix and an affine gap cost of $11 + k$ for each gap of length $k$. While the lower degree of sequence similarity in the PDB40 database compared to the PDB90 database leads to a significant difference in sensitivity between the two sequence sets, comparison with results of the Smith–Waterman algorithm using the same scoring functions show that the performance of the two algorithms on the same test set is hardly distinguishable. For a position-specific scoring function (taken from the Pfam ig model) the sensitivity approaches the maximum of 1.0 expected of the HMMer program used to build the ig model.

| Test set<br>(gold standard) | PDB40D-B<br>(SCOP) | PDB90D-B<br>(SCOP) | Pfam A<br>(ig) |
|---|---|---|---|
| Hybrid | 0.791 | 0.855 | 0.992 |
| S-W | 0.767 | 0.844 | — |

the sensitivity measures summarized in Table 1 indicate quite clearly that differences in the algorithm play only a minor role, and the degree of sensitivity is determined more critically by the choices of the scoring function, test set, and gold standard. Consequently, we observe that the much simpler significance assessment of the hybrid algorithm comes at no significant penalty in performance, which is the main result of this study.

## 4 SUMMARY AND OUTLOOK

We have established in this study that the null statistics of the hybrid alignment scores obey the Gumbel distribution with the Gumbel parameter $\lambda = 1$ for a large class of scoring functions, including the more than 2000 models of protein domains contained in Pfam. Furthermore, we have presented a performance evaluation composed of a thorough study of the position-independent case and preliminary results for position-specific scoring systems. It indicates that the sensitivity of the hybrid algorithm is comparable to the best of the existing algorithms when the same scoring functions are used. Therefore, hybrid alignment is a high-performance alignment algorithm with well-characterized, universal score statistics. The latter eliminates the time-consuming simulations normally needed to assign $p$-values to alignment scores. In fact, even the small finite-length corrections to the score statistics can be computed very efficiently using methods described in Appendix B and by Yu and Hwa (2001).

Since the hybrid algorithm (as presented in detail in Appendix A) is a combination of the Smith–Waterman and the probabilistic local alignment algorithm, its computational complexity is the same as the latter two, i.e. the computational time of each alignment scales as the product of the lengths of the two sequences. This is too slow for large database searches. However, it can be sped up tremendously by applying the BLAST heuristics, which looks for 'diagonals' of high scores, i.e. runs of high scores without gaps, and performs the full gapped alignment only for those cases which contain promising diagonals. These heuristics will reduce the computational time to the sum of the two sequence lengths as in BLAST. One interesting (maybe also important) feature of a database search tool based on hybrid alignment is that the universal statistics of hybrid alignment allows automatic adjustment for unusual amino acid composition of individual query sequences in the assignment of $p$-values, since the theory of Yu and Hwa (2001) does allow different amino acid compositions for the two sequences being compared. Also, hybrid alignment allows an interactive search for good scoring parameters since it is not limited to scoring parameters for which the Gumbel parameters have been precomputed by simulations. A preliminary implementation of the hybrid algorithm with BLAST heuristics for this purpose is available upon request from the authors.

The ability of the hybrid alignment to provide instantaneous and accurate score statistics should be most useful to applications which iteratively fine-tune the scoring functions to improve alignment sensitivity. A well-known example is the PSI-BLAST program, which fine-tunes the scoring functions each round according to the 'significant' sequences extracted from alignments in the previous round. This is a case which explicitly requires the evaluation of $p$-values for different scoring functions in real time. The universal score statistics of hybrid alignment will enable PSI-BLAST to accommodate position-specific gap functions for the first time, with a potential to significantly improve its sensitivity for remote homology detection.

## ACKNOWLEDGEMENTS

YKY wishes to thank Shayn-Ming Perng, Chien-Jung Huang, and in particular Hungwen Chang for useful correspondence. We also benefited from discussions with Stephen Altschul, Ewan Birney, and Rolf Olsen. YKY is grateful to the hospitality of the University of California, San Diego, where part of this study was carried out. RB and TH acknowledge the financial support by the Beckman Foundation and the NSF through Grant No. DMR-9971456.

## REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Altschul,S.F., Madden,T.L., Schäffer,A.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.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam contribution to the annual NAR database issue. *Nucleic Acids Res.*, **28**, 263–266.

Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A.G. (1996) Understanding protein structure: using SCOP for fold interpretation. *Meth. Enzymol.*, **266**, 635–643.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Bucher,P. and Hofmann,K. (1996) A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States,D.J. *et al.* (ed.), *Proceedings of The Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB96)*. AAAI Press, Menlo Park, pp. 44–50.

Bundschuh,R. (2000) An analytic approach to significance assessment in local sequence alignment with Gaps. In Istrail,S. *et al.* (ed.), *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*. ACM press., New York, pp. 86–95.

Collins,J.F., Coulson,A.F.W. and Lyall,A. (1988) The significance of protein sequence similarities. *CABIOS*, **4**, 67–71.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. and Eck,R.V. (eds), *Atlas of Protein Sequence and Structure*, 5 Suppl., **3**, pp. 345–358. Natl Biomed. Res. Found.

Eddy,S., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden markov models of sequence consensus. *J. Comp. Biol.*, **2**, 9–23.

Eddy,S. (1995) Multiple alignment using hidden markov models. In Rawlings,C. *et al.* (ed.), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, pp. 114–120.

Eddy,S.R. (1997) Maximum likelihood fitting of extreme value distributions. Unpublished technical notes. Available at `http://www.genetics.wustl.edu/eddy/publications/`.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Karlin,S. and 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.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Mott,R. (1992) Maximum likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bull. Math. Biol.*, **54**, 59–75.

Mott,R. and Tribe,R. (1999) Approximate statistics of gapped alignment. *J. Comp. Biol.*, **6**, 91–112.

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.

Olsen,R., Bundschuh,R. and Hwa,T. (1999) Rapid assessment of extremal statistics for gapped local alignment. In Lengauer,T. *et al.* (ed.), *Proceedings of The Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB99)*. AAAI Press, Menlo Park, pp. 211–222.

Pearson,W.R. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Siegmund,D. and Yakir,B. (2000) Approximate P-values for sequence alignments. *Ann. Stat.*, **28**, 657–680.

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

Smith,T.F., Waterman,M.S. and Burks,C. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.*, **13**, 645–656.

Waterman,M.S. and Vingron,M. (1994a) Sequence comparison significance and poisson approximation. *Stat. Sci.*, **9**, 367–381.

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

Yu,Y.-K. and Hwa,T. (2001) Statistical significance of probabilistic sequence alignment and related local hidden markov models. *J. Comp. Biol.*, **8**, 249–282.

## Appendix A: The hybrid alignment algorithm

The hybrid alignment algorithm developed by Yu and Hwa (2001) for uniform scoring functions is first reviewed in this Appendix, followed by the algorithm for position-specific scoring functions. Let the two protein sequences being aligned be $\mathbf{a} = [a_1, a_2. \cdots a_M]$ and $\mathbf{b} = [b_1, b_2, \cdots, b_N]$, where $a_m$ is the $m^{\text{th}}$ amino acid of sequence $\mathbf{a}$, and $b_n$ is the $n^{\text{th}}$ amino acid of sequence $\mathbf{b}$. The basic algorithm for uniform substitution and affine gap cost consists of recursive iteration of the following equations,

$$\begin{aligned} Z^S_{m,n} &= 1 + \eta\ W(a_m, b_n) \cdot [Z^S_{m-1,n-1} \\ &\qquad + \mu^{D1}\ Z^D_{m-1,n-1} + \mu^{I1}\ Z^I_{m-1,n-1}], \\ Z^D_{m,n} &= \mu^{D2}\ Z^S_{m-1,n} + \nu^D\ Z^D_{m-1,n}, \\ Z^I_{m,n} &= \mu^{I2}\ Z^S_{m,n-1} + \nu^I\ Z^I_{m,n-1} + \mu' \mu^{I2} \mu^{D1}\ Z^D_{m,n-1} \end{aligned} \tag{2}$$

for the three auxiliary quantities $Z^S_{m,n}$, $Z^D_{m,n}$ and $Z^I_{m,n}$, from $m = 0$ to $M$ and $n = 0$ to $N$, with the boundary conditions

$$\begin{aligned} &Z^D_{m\le 0,n\ge 0} = 0,\ Z^D_{m\ge 0,n<0} = 0, \\ &Z^I_{m<0,n\ge 0} = 0,\ Z^I_{m\ge 0,n\le 0} = 0, \\ &Z^S_{m<0,n\ge 0} = 0,\ Z^S_{m\ge 0,n<0} = 0, \\ &Z^S_{m=0,n\ge 0} = 1,\ Z^S_{m\ge 0,n=0} = 1. \end{aligned} \tag{3}$$

The alignment score $\mathsf{S}$ is computed as

$$\mathsf{S}[\mathbf{a}, \mathbf{b}] = \max_{\substack{1 \le m \le M \\ 1 \le n \le N}} \{\ln Z_{m,n}\}, \tag{4}$$

where

$$Z_{m,n} = Z^S_{m,n} + Z^D_{m,n} + Z^I_{m,n}. \tag{5}$$

We remark that Equation (4) together with the Viterbi version of Equations (2) and (5) turns the algorithm to the Smith–Waterman algorithm. On the other hand, taking Equations (2) and (5) as they are, but replacing Equation (4) by $e^{\mathsf{S}} = \sum_{m,n} Z_{m,n}$ turns the algorithm into a simplified version of the probabilistic local alignment introduced by Bucher and Hofmann (1996). The algorithm specified by Equations (2), (4) and (5) has just the appropriate mixture of the Smith–Waterman and probabilistic local alignment (hence the name 'hybrid') that it produces universal statistics for the alignment score $\mathsf{S}$, as long as the parameters of Equation (2) satisfy some weak condition as specified below.

The possible input parameters to Equation (2) are the substitution matrix $W(a, b)$, the conservation parameter $\eta$, and the affine gap parameters $\mu^{D1}$, $\mu^{D2}$, $\mu^{I1}$, $\mu^{I2}$, $\nu^D$, $\nu^I$. There is one additional parameter $\mu'$ which is either set to 1 if a double-gap (i.e. an insertion immediately following a deletion, or a deletion immediately following an insertion) is allowed, or is set to 0 if the double gaps are not allowed.

As described in Yu and Hwa (2001), the hybrid algorithm is defined in a certain subspace of the above scoring parameter space. These parameters can be represented more succinctly in terms of the usual input to a Smith–Waterman-type algorithm, which contains a substitution score $s(a, b)$ (e.g. the PAM or BLOSUM scoring matrix) and an affine gap cost, say $d + \epsilon k$ for each gap of length $k$. In terms of $s(a, b)$, $d$ and $\epsilon$, the parameters become

$$W(a, b) = \exp[\lambda_{\text{ug}} s(a, b)] \tag{6}$$

where $\lambda_{\text{ug}}$ is the unique positive root of the equation

$$\sum_{a,b} \mathrm{e}^{\lambda_{\text{ug}} s(a,b)} p(a) p(b) = 1, \tag{7}$$

for a given model of amino acid background frequency $p(a)$. The other 7 parameters in Equation (2) are defined in terms of

$$\mu = \exp[-\lambda_{\text{ug}}(d + \epsilon)] \tag{8}$$

$$\nu = \exp[-\lambda_{\text{ug}}\epsilon] \tag{9}$$

and $\mu' \in \{0, 1\}$ as

$$\eta = (1 - \nu)^2 / [(1 + \mu - \nu)^2 + (\mu' - 1)\mu^2], \tag{10}$$

$$\mu^{I1} = [(1 + \mu - \nu)^2 + (\mu' - 1)\mu^2] / (1 - \nu), \tag{11}$$

$$\mu^{D1} = [(1 + \mu - \nu)^2 + (\mu' - 1)\mu^2] / (1 + \mu'\mu - \nu), \tag{12}$$

$$\mu^{I2} = \mu(1 - \nu) / [(1 + \mu - \nu)^2 + (\mu' - 1)\mu^2], \tag{13}$$

$$\mu^{D2} = \mu(1 + \mu'\mu - \nu) / [(1 + \mu - \nu)^2 + (\mu' - 1)\mu^2], \tag{14}$$

$$\nu^D = \nu, \tag{15}$$

$$\nu^I = \nu. \tag{16}$$

The theory of Yu and Hwa (2001) predicts that the alignment score $\mathsf{S}$ obtained using Equations (2)–(5) with the parameters chosen according to Equations (6)–(16) satisfies the Gumbel distribution with the key Gumbel parameter $\lambda = 1$ for arbitrary[¶] scoring functions $s(a, b)$, $d$, and $\epsilon$ in the asymptotic limit $M \gg 1$ and $N \gg 1$. The performance evaluation of the hybrid algorithm on the SCOP sequences described in Section 3 was done with $s(a, b)$ being the BLOSUM62 substitution scores, $d = 11$, $\epsilon = 1$ and $\mu' = 1$. The study of the Smith–Waterman algorithm in Section 3 was done with the Viterbi version of Equations (2) and (5), with the same BLOSUM62 substitution scores, $d = 11$, $\epsilon = 1$, $\mu' = 1$, and also with $\eta = 1$.

For the general position-specific scoring functions, all the alignment parameters become position-dependent; the insertion/deletion weights generally also become amino-acid dependent. Equation (2) is now replaced by

$$\begin{aligned} Z^S_{m,n} &= 1 + \eta_{m-1,n-1}\, W_{m,n}(a_m, b_n) \cdot [Z^S_{m-1,n-1} \\ &\quad + \mu^{D1}_{m-1,n-1}\, Z^D_{m-1,n-1} \\ &\quad + \mu^{I1}_{m-1,n-1}\, Z^I_{m-1,n-1}], \\ Z^D_{m,n} &= \mu^{D2}_{m-1,n}(a_{m-1})\, Z^S_{m-1,n} + \nu^D_{m-1,n}(a_{m-1})\, Z^D_{m-1,n}, \\ Z^I_{m,n} &= \mu^{I2}_{m,n-1}(b_n)\, Z^S_{m,n-1} + \nu^I_{m,n-1}(b_n)\, Z^I_{m,n-1} \\ &\quad + \mu' \mu^{I2}_{m,n-1}(b_n) \mu^{D1}_{m,n-1}\, Z^D_{m,n-1}, \end{aligned} \tag{17}$$

with the boundary conditions still given by Equation (3). In Section 3, we apply the above algorithm to the class of Pfam models. Each Pfam model of protein domain is a hidden Markov model (HMM) consisted of a linear arrangement of 'nodes'. The total number of nodes $M$ is called the 'model length'. Associated with each node $m$, there are 9 transition probabilities $P_m(X_m \to X_{m+1})$, where $X \in \{S, I, D\}$ represents one of the 3 possible 'hidden states', with '$S$' for substitution, '$I$' for insertion, and '$D$' for deletion. In the '$S$' and the '$I$' states, the model can 'emit' amino acid $b$ with frequencies $S_m(b)$, and $I_m(b)$ respectively. The Pfam models do not allow for direct transition between the '$I$' and '$D$' states, hence, $P_m(I \to D) = 0$ and $P_m(D \to I) = 0$ for all $m$'s. The models are complemented by the boundary conditions

[¶] The substitution score $s(a, b)$ needs to have at least one positive entry.

**Table 2.** Correspondence between the Pfam HMM model parameters and the hybrid alignment parameters

| Hybrid | Pfam HMM |
|---|---|
| $W_{m,n}(a,b)$ | $S_m(b)$ |
| $\eta_{m\geq 1,n\geq 0}$ | $P_m(S \to S)$ |
| $\mu^{I2}_{m\geq 1,n\geq 0}(b_{n+1})$ | $I_m(b_{n+1})P_m(S \to I)$ |
| $\mu^{D2}_{m\geq 1,n\geq 0}(a_m)$ | $P_m(S \to D)$ |
| $\mu^{I1}_{m\geq 1,n\geq 0}$ | $P_m(I \to S)$ |
| $\mu^{D1}_{m\geq 1,n\geq 0}$ | $P_m(D \to S)$ |
| $\nu^{I}_{m\geq 1,n\geq 0}(b_{n+1})$ | $I_m(b_{n+1})P_m(I \to I)$ |
| $\nu^{D}_{m\geq 1,n\geq 0}(a_m)$ | $P_m(D \to D)$ |
| $\mu'$ | 0 |
| $\eta_{m=0,n\geq 0}$ | $P(I_N \to B)P(B \to S)$ |
| $\mu^{I2}_{m=0,n\geq 0}(b)$ | $P(I_N \to I_N)$ |
| $\mu^{D2}_{m=0,n\geq 0}(a)$ | $P(I_N \to B)P(B \to D)$ |
| $\mu^{I1}_{m=0,n\geq 0}$ | 1 |
| $\mu^{D1}_{m=0,n\geq 0}$ | 0 |
| $\nu^{D}_{m=0,n\geq 0}$ | 1 |
| $\nu^{I}_{m=0,n\geq 0}$ | $P(I_N \to I_N) + P(I_N \to B)P(B \to D)$ |

that allow null insertions before getting into the begin state $B$ with probability $P(I_N \to I_N) = 1 - P(I_N \to B)$. The begin state can go to the $S$ state or the $D$ state with probabilities $P(B \to S)$ and $P(B \to D)$ respectively.

A Pfam model can be easily accommodated in the alignment algorithm Equation (17), when the model is viewed as the sequence **a**. The Pfam parameters correspond to some special choices of the alignment parameters in Equation (17) as indicated in Table 2. The score statistics and performance of the Pfam models reported in Sections 2 and 3 are obtained by iterating Equation (17), augmented by Equations (3)–(5), with alignment parameters specified according to Table 2 for each of the Pfam models. As shown in the text, the alignment score $\mathsf{S}$ obeys the Gumbel distribution with $\lambda = 1$ (up to small length-dependent corrections described in (Yu and Hwa, 2001)) for all of the more than 2000 Pfam models tested.

## Appendix B: The Gumbel parameter $K$

Our theory does not yet provide us with the value of the other Gumbel parameter $K$. However, the parameter $K$ only fixes the position of the center of the Gumbel distribution, i.e. the $p$-value of typical alignment scores. Thus, it plays a much less significant role in the assignment of $p$-values than the parameter $\lambda$. (The latter describes the tail of the distribution and therefore the interesting regime of low $p$-values.)

Furthermore, $K$ can be determined very rapidly numerically once $\lambda$ is known. One may, for example, perform 50 pairwise alignments and convert the average score into an estimate of $K$ via the expression

$$\langle \mathsf{S} \rangle_0 = [\ln KMN + \gamma]/\lambda \tag{18}$$

for the expectation value of a Gumbel distributed variable $\mathsf{S}$ where $\gamma = 0.5772156..$ is Euler's constant.