Using Hybrid Alignment for Iterative Sequence Database Searches

Yuheng Li, Mario Lauria,
Department of Computer and Information Science
The Ohio State University
2015 Neil Avenue #395
Columbus, OH 43210-1106 U.S.A.
lauria@cis.ohio-state.edu

Ralf Bundschuh
Department of Physics
The Ohio State University
174 West 18th Avenue
Columbus, OH 43210-1106 U.S.A.
bundschuh@mps.ohio-state.edu

Abstract

Progressive sequence model refinement by means of iterative searches is an effective technique for high sensitivity database searches and is currently employed in popular tools such as PSI-BLAST and SAM. Recently, a novel alignment algorithm has been proposed that offers features expected to improve the sensitivity of such iterative approaches, specifically a well-characterized theory of its statistics even in the presence of position-specific gap costs. Here, we demonstrate that the new hybrid alignment algorithm is ready to be used as the alignment core of PSI-BLAST. In addition, we evaluate the accuracy of two proposed approaches to edge effect correction in short sequence alignment statistics that turns out to be one of the crucial issues in developing a hybrid-alignment based version of PSI-BLAST.

1 Introduction

The availability of large numbers of entire genomes requires powerful bioinformatics tools to assign meaning to the sequence data and to leap forward into areas like proteomics. Perhaps the most fundamental and widely used tool of genomic analysis is sequence alignment. Although sequence alignment is a well-established technique, the need to detect weaker and weaker sequence homologies requires continuous improvements in the sensitivity of alignment algorithms.

The most commonly used sequence alignment tools like BLAST [3] and FASTA [25] are based on the Smith-Waterman algorithm [30]. Recently, a variation of the Smith-Waterman algorithm called hybrid alignment has been proposed [36]. The main advantage of hybrid alignment is that it is backed by a theory of its statistics that allows to quickly assign reliable E-values for arbitrary scoring systems. Handling of arbitrary scoring systems is particularly relevant in iterative algorithms like PSI-BLAST [3] or SAM [16] that dynamically adapt their scoring system to the query sequence. Such iterative approaches are required to detect more remote homologies than those found by BLAST or FASTA.

Hybrid alignment has the same computational complexity as the Smith-Waterman algorithm. However, it has also been combined with the heuristic approaches of BLAST to render it computationally efficient [35]. It has been shown to be comparable to the Smith-Waterman algorithm in its ability to detect sequence homologies in pairwise sequence alignments [35]. However, it has never been evaluated within the iterative framework that it is presumably most useful for.

Here, we put the hybrid algorithm to test within an established iterative search framework, namely PSI-BLAST. In principle, the hybrid algorithm can offer crucial additional features to PSI-BLAST such as position-specific gap costs. However, the heuristics built into a high performance tool like PSI-BLAST have been extensively optimized for its native, Smith-Waterman based, alignment algorithm over the course of several years. Thus, it is not clear what would be the effects of replacing the current algorithm for E-value computation with the hybrid algorithm, even without adding extra features such as position specific gap costs. For example, we found that the edge effect correction of E-values follows different laws between the two versions of PSI-BLAST, pointing to at least one area in which the two algorithms (and their underlying statistics) interact differently with the rest of the code.

The purpose of this study is to answer the question of the hybrid algorithm's compatibility as an E-value computation engine in PSI-BLAST. Assessing this compatibility is crucial in determining the extent to which the hybrid algorithm can leverage the investments that went into building the current crop of high performance bioinformatics tools.

This paper offers two main contributions. The first is to show that the hybrid alignment algorithm is ready to be used used as the alignment core of PSI-BLAST, and that only minimal change is required to PSI-BLAST (the edge effect correction formula). The second main contribution is to evaluate the accuracy of two proposed approaches to edge effect correction in short sequence alignment statistics.

The remainder of the manuscript is organized as follows. Section 2 provides some background on sequence alignment statistics. Section 3 presents the steps necessary to incorporate hybrid alignment as the alignment core of PSI-BLAST. Then, Section 4 describes in detail how sequence length effects are being taken into account since this differs significantly from the original version of PSI-BLAST. The direct comparison of the hybrid and the original version of PSI-BLAST is presented in Section 5. Finally, Section 6 concludes the paper.

2 Review of sequence alignment statistics

Pairwise sequence alignment algorithms assign an alignment score to each pair of sequences. The score is the larger the more similar the two sequences are. Iterative sequence alignment tools like PSI-BLAST or SAM build up on these pairwise sequence alignment algorithms. In each iteration the pairwise sequence alignment algorithm is used to search a large sequence database which leads to a list of hits ordered by their score. From this list of hits a multiple alignment is created that in turn determines the scoring system of the next iteration. The crucial step from one iteration to the other is the decision which of the hits to keep as putative members of the family (and include in the multiple alignment) and which of the hits to reject as non-relevant. A reliable quantitative criterion for which sequences to keep as putative members and which to discard as random hits is a cutoff in the E- or p-value of the sequences.

The statistical significance expressed by the E-value judges the quality of an alignment relative to all alignments that one would obtain by aligning randomly chosen (and thus unrelated) sequences. Thus, it can only be calculated if it is known how the alignment scores of randomly chosen sequences are distributed. For alignment algorithms that do not allow gaps, i.e., insertions or deletions, in their alignments this alignment score distribution of random sequences is known. It has been rigorously proven [14, 15, 12] that the expected number of gapless local alignments of two sequences of length M and N with a score larger that Σ, i.e., the E-value, follows in the limit of infinitely long sequences the universal form

$$E(\Sigma) = KMNe^{-\lambda\Sigma}. \qquad (1)$$

This form neither depends on the scoring parameters nor on the sequence model, i.e., the frequencies with which each amino acid appears in the random sequences, as long as the parameters are chosen such that the alignments are really local. However, the two parameters λ and K do depend on the scoring parameters. The Karlin-Altschul theory also describes this dependence. Thus, an E-value can be assigned to a gapless alignment without any further need for computation which made the original version of BLAST so successful.

However, in order to detect weak sequence homologies, it is crucial to allow gaps in an alignment [24]. In the presence of gaps the E-values follow according to numerical studies still the universal form Eq. (1) [31, 10, 19, 33, 34, 2, 23] However, the numerical values of the two parameters λ and K are not known.

There are various approaches to solve this dilemma: for large gap costs there are approximate analytical formulas for λ [21, 20, 29] For a small sub-class of scoring systems there is even an analytical formula for λ that is valid for all gap costs [6]. The current version of PSI-BLAST uses a heuristic method to estimate λ for different scoring matrices but at fixed gap cost [3, 28, 27] and there are numerical approaches [7, 8] as well to rapidly determine λ.

However, all of these approaches are either heuristic or restricted to certain regimes of the alignment parameters. A possible escape route from this dilemma is an alternative alignment algorithm that has been proposed by Yu and Hwa [36]. The algorithm is called hybrid alignment since it is a combination of the Smith-Waterman algorithm and probabilistic schemes like hidden Markov models. In hybrid alignment the Smith-Waterman algorithm is modified such that its E-values are still calculated according to Eq. (1) but with the parameter λ taking the universal value $\lambda = 1$ completely independently of the scoring system. This simplification of the statistics does not decrease the sensitivity of the algorithm compared to the traditional Smith-Waterman algorithm [35]. The basic computational complexity of the alternative algorithm is the same as for Smith-Waterman and it can be combined with heuristic schemes similar to the ones used in BLAST to reduce the computational effort. Most importantly, the theoretical prediction of the universal form Eq. (1) with $\lambda = 1$ holds even for position-dependent gap costs. This prediction has also been numerically verified [35] for a large range of scoring systems with position dependent gap costs taken from the PFAM [4] database. The inability to calculate E-values for position-specific gap costs is precisely the reason why PSI-BLAST does not incorporate a position-specific gap cost in spite of the expectation that such a position-specific gap cost would increase sensitivity significantly if it were possible to implement it. Thus, using hybrid alignment in PSI-BLAST would not only provide a theoretical basis for the calculation of E-values with the current fixed gap cost scoring systems but also open up the possibility to the future incorporation of

more sensitive position-specific gap costs.

3 Incorporation of hybrid alignment into PSI-BLAST

As part of a previous work, a non-position-specific version of the hybrid alignment algorithm was incorporated into version 2.0 of the freely available NCBI BLAST source code [35]. The hybrid version of BLAST (HYBLAST) retains the familiar user interface and many of the features of the original NCBI BLAST. It uses the same heuristics for deciding which database sequence is a potential hit that gives the original BLAST its huge speed advantage over full Smith-Waterman. However, by replacing the alignment core of BLAST by hybrid alignment the modified version is capable to deal with any scoring system and gap cost the user wishes to provide. Due to the more involved statistics of the Smith-Waterman algorithm the original BLAST forces the user to choose a combination of a substitution matrix and gap costs from a preselected set for which the statistics has been pre-calculated in time-consuming computer simulations.

Since PSI-BLAST is an extension of BLAST we took the HYBLAST program as our starting point for the implementation of the hybrid algorithm in PSI-BLAST. Due to the similarity of their user interfaces, only minimal changes were required to adapt the core components of PSI-BLAST to work with HYBLAST instead of BLAST as described below. Therefore the results of our comparative measurements can be attributed purely to the differences in the statistics underlying the two algorithms and the way they interact with PSI-BLAST heuristics, and not to code dissimilarities, as required to fulfill the objectives of our study.

First, the alignment routines themselves had to be changed such that they use the position-specific weight matrix instead of the uniform scoring matrix used in BLAST and HYBLAST. The new routines implement the recursion equations for position-specific hybrid alignments given by Yu, Bundschuh, and Hwa [35].

Second, the position-specific weight matrix has to be filled during the model building phase of PSI-BLAST. The original PSI-BLAST code calculates for each position i of the query sequence the 20 probabilities $p_{i,a}$ to observe amino acid a at this position. These probabilities are derived from the actually observed amino acids at this position and from prior expectations determined by the amino acid in the query sequence in case that there are only very few sequences in the multiple alignment. The position-specific scoring matrix $s_{i,a}$ of PSI-BLAST at position i for amino acid a is then assigned a score of $s_{i,a} = \log(p_{i,a}/p_a)$ where p_a is the background probability to observe the amino acid a. Afterwards these scores are rescaled in some particular way [3]. Since the position-specific alignment weight used by the hybrid algorithm is simply $p_{i,a}/p_a$ itself, the position-specific alignment weight matrix can easily be filled together with the usual position-specific score matrix of PSI-BLAST. In contrast to the scoring matrix the weight matrix does not require any rescaling.

We did not implement position specific gap costs for our experiments. While highly desirable, the implementation of such feature would require non trivial changes to the existing PSI-BLAST code, and it is beyond the scope of this paper.

After creating a version of PSI-BLAST incorporating the hybrid sequence alignment algorithm (in the following called Hybrid PSI-BLAST), we performed a series of measurements to compare its performance to that of the unmodified version of PSI-BLAST 2.0 (called NCBI PSI-BLAST in the following). In the course of these comparisons we realized that the performance of Hybrid PSI-BLAST was influenced by the edge effect correction formula being employed. The next section describes the approach we used to determine the correct formula to implement, selected among those found in the literature. In the ensuing section we report the results of the Hybrid vs. NCBI comparison, with the Hybrid incorporating the formula so determined.

4 Edge-effect correction

Eq. (1) is only valid in the limit of infinitely long sequences. Since sequences in database searches can be rather short, it is important to correct E-values for the finite sequence length. For this purpose two different correction formulas have been suggested. Both involve in addition to the parameters K and λ of Eq. (1) the relative entropy H and the offset β of the scoring system. These quantities depend on the specific scoring system and have to be determined numerically.

Altschul and Gish [2] proposed a formula that has later been extended by Altschul, Bundschuh, Olsen, and Hwa [1] to read for a sequence pair of length N and M, respectively:

$$E(\Sigma) = K\left[N - \left(\frac{\lambda\Sigma}{H} + \beta\right)\right]\left[M - \left(\frac{\lambda\Sigma}{H} + \beta\right)\right] e^{-\lambda\Sigma} \quad (2)$$

The alternative formula used by Yu and Hwa [36] is

$$E(\Sigma) = K(N-\beta)(M-\beta)\times \quad (3)$$
$$\times \exp\left[-\lambda\left\{1 + \frac{1}{(M-\beta)H} + \frac{1}{(N-\beta)H}\right\}\Sigma\right].$$

Analytical approaches to the edge correction problem [18, 32] are confined to alignment without gaps. Even in the absence of gaps, they only give corrections of Eq. (1) to first order in $\lambda\Sigma/[(N-\beta)H]$. Both correction formulas Eqs. (2) and (3) coincide up to first order in $\lambda\Sigma/[(N-\beta)H]$. Thus, the analytical results are not suited to distinguish one

correction formula from the other even in the absence of gaps.

The equivalence of the two correction formulas up to terms of order $\lambda\Sigma/[(N-\beta)H]$ is also the reason why the existence of different formulas was not an issue for the conventional PSI-BLAST. For the default scoring system of PSI-BLAST, i.e., the BLOSUM62 scoring matrix [13] with cost of $11+k$ for a gap of length k and the amino acid frequencies of Robinson and Robinson [26], the parameters are estimated to be $\lambda \approx 0.2670$, $K \approx 0.042$, $H \approx 0.14$, and $\beta \approx -30$ [1]. At a database size of $M = 10^6$ amino acids and a query size of $N = 100$ amino acids an E-value of one corresponds to a score of $\lambda\Sigma \approx 15$. Thus, the first order correction is $\lambda\Sigma/[(N-\beta)H] \approx 0.77$ which is sizeable but still smaller than one (i.e., the correction is smaller than the leading term and higher order corrections are expected to become even smaller.)

The situation in hybrid alignment is different. For the same scoring system, the parameters are estimated as $\lambda = 1$, $K \approx 0.3$, $H \approx 0.07$, and $\beta \approx -50$. The larger value of K implies that an E-value of one now corresponds to a score of $\lambda\Sigma \approx 17$. More importantly, due to the smaller value of the relative entropy H, the first order correction is $\lambda\Sigma/[(N-\beta)H] \approx 1.6 > 1$. Thus, the second order correction contributes significantly to the E-value. Therefore, it has to be determined for hybrid alignment which of the two formulas more appropriately describes the length dependence of the E-values or if yet another formula has to be worked out.

We address this point empirically by aligning sequences from a database derived from SCOP [22, 11] by the astral compendium [17, 9] (http://astral.stanford.edu/, release ASTRAL SCOP 1.59). We use a database that contains only sequences with less than 40% pairwise sequence identity. We use every sequence from the database as a query for a hybrid alignment search of the whole database. This yields a list of hits for each query and their respective E-values calculated by formula Eq. (2) or (3). Following the approach of BLAST and PSI-BLAST instead of evaluating Eqs. (2) and (3) for each hit, we use

$$E(\Sigma) = K A_{\text{eff}} e^{-\lambda\Sigma} \tag{4}$$

where A_{eff} is the effective search space. It is determined once for each query as

$$A_{\text{eff}} \equiv \frac{e^{\lambda\Sigma^*}}{K} \tag{5}$$

with Σ^* given by $E(\Sigma^*) = 1$ according to Eq. (2) or (3). In this framework the difference between Eqs. (2) and (3) translates into a different value of Σ^*, i.e., in a different value of the effective search space A_{eff}.

Since the database is derived from the structural SCOP database, hits can be identified as true homologs if they belong to the same superfamily in SCOP and as non-homologs if not. Thus, for each E-value cutoff the number of errors per query can be calculated as the number of non-homologs

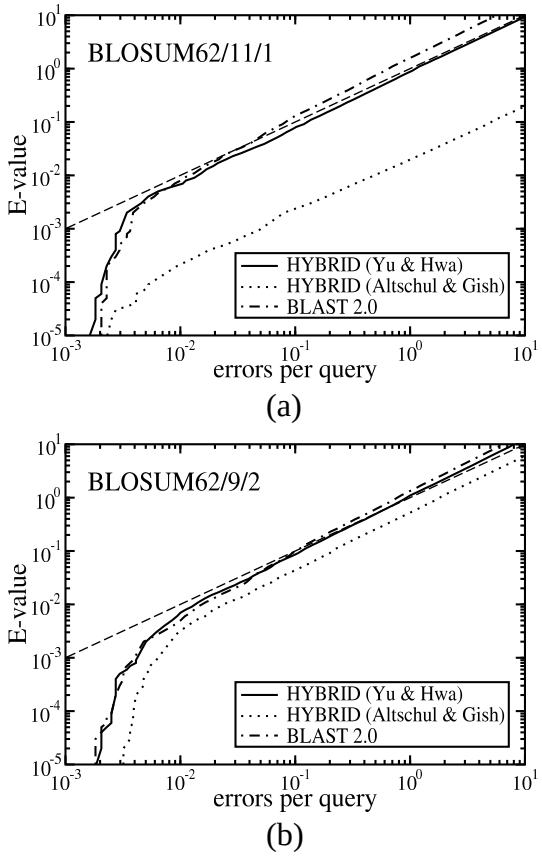

Figure 1. Comparison of two formulas for edge effect correction. Both graphs show the dependence between the E-value cutoff and the number of errors per query, i.e., the number of non-homologous sequence pairs with an E-value lower than the cutoff divided by the total number of sequences in the database. The dotted line corresponds to hybrid alignment with E-values calculated according to Eq. (2) while the solid line corresponds to hybrid alignment with E-values calculated according to Eq. (3). The dash-dotted line is the result of BLAST 2.0 and the dashed line is the identity corresponding to an ideal algorithm. Both graphs show data for the ASTRAL40 database and the BLOSUM62 scoring matrix. In (a) the cost of a gap of length k is $11+k$ while in (b) it is $9+2k$. In both cases BLAST 2.0 and Eq. (3) yield good estimates of the E-value while Eq. (2) is clearly inferior for hybrid alignment.

with an E-value lower than the given cutoff divided by the total number of queries in the dataset, which is $4{,}383$ in this case. If the calculation of E-values is correct, the number of errors per query is identical to the E-value cutoff.

Figure 1 shows the relationship between the errors per query and the E-value cutoff for the PSI-BLAST default scoring system and for the BLOSUM62 scoring matrix with a cost of $9 + 2k$ for a gap of length k. The latter has a relative entropy of $H \approx 0.15$ and thus the contribution of the higher order terms should be less dramatic than for the PSI-BLAST default scoring system. Each of the two graphs shows the plots of the E-value versus the errors per query for hybrid alignments with both length correction formulas and of the original BLAST 2.0. The (ideal) identity is shown as the dashed line. In both cases it can be seen that the E-values obtained by Eq. (3) are very close to ideal. The E-values of BLAST 2.0 are similarly good while the E-values for the hybrid algorithm calculated according to Eq. (2) are too small. As expected based on the differences in the relative entropies the effect is much stronger for the BLOSUM62/11/1 scoring system than for the BLOSUM62/9/2 scoring system. We conclude that for the hybrid alignment algorithm Eq. (3) provides good estimates of the E-value while Eq. (2) should not be used.

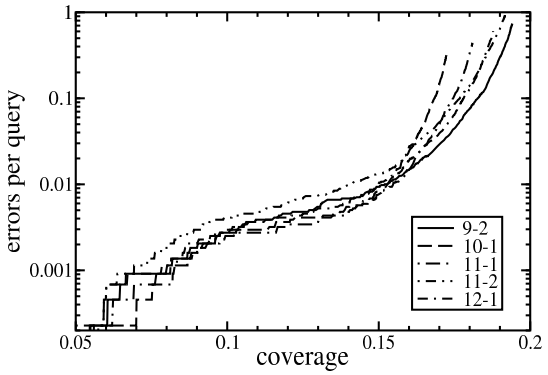

Figure 2. Comparison of Hybrid PSI-BLAST performance for different gap costs. The curves show the trade-off between the errors per query and the coverage for Hybrid PSI-BLAST on a "gold standard" database using different gap costs. While all curves are relatively close together, a cost of $11 + 1 \cdot k$ for a gap of length k seems to lead to the best performance.

5 Hybrid versus NCBI comparison

For the Hybrid version of PSI-BLAST the way of calculating effective query length, effective database length and effective search space was modified as described in the previous section to be able to deal with small values of the relative entropy H. In the NCBI version of PSI-BLAST, the value H is looked up from a table and is not very small, but in the hybrid version H is calculated and takes on small values for some queries and/or scoring matrices.

In order to study the performance of our prototype Hybrid PSI-BLAST we applied the alignment sensitivity assessment of Brenner, Chothia and Hubbard [5]. The same SCOP derived dataset as described in the last section was used [17, 9] (http://astral.stanford.edu/, release ASTRAL SCOP 1.59). However, since we suspected a possible true relationship not reflected in the SCOP classification we removed a single sequence that was consistently misclassified by all versions (Hybrid and NCBI) of the algorithm for nearly all parameter choices, namely the representative of the superfamily c.11.1.

Using this database as a "gold standard" we performed two different sensitivity assessments. First, we used each of the sequences in the gold standard database as queries. For each query we searched the gold standard database with the Hybrid and the NCBI version of PSI-BLAST. We ran both PSI-BLAST versions for several iterations until they converged. From the resulting lists of hits with their E-values we calculated for each E-value cutoff in addition to the errors per query described in the last section the coverage, i.e., the number of true hits with an E-value smaller than the cutoff divided by the total number of true hits in the database which is $88{,}171$ in our study. The plot of errors per query versus coverage as the E-value cutoff is varied represents the relationship between the sensitivity and selectivity of a program.

In this test we observed a large increase in computational effort when using the HYBRID algorithm. The total computer time required for the assessment of the HYBRID algorithm was about ten times higher than for the original PSI-BLAST. However, this is an artefact of the unrealistically small database size in this test. The HYBRID algorithm requires some query-dependent parameters like the relative entropy H to be calculated during the startup phase. For a short database this startup phase dominates the computational effort. For longer databases, the computational effort for the startup phase is not important any more and the computational effort of the HYBRID algorithm and PSI-BLAST become comparable (see below.)

Since the hybrid algorithm treats gaps differently from the Smith-Waterman algorithm underlying the NCBI PSI-BLAST, it is not a priori clear if the cost of $11 + 1 \cdot k$ for a gap of length k that has been determined to be optimal for the original PSI-BLAST is also good for the hybrid version. Thus, we first compared different values of the gap initiation and extension cost for the hybrid version of PSI-BLAST given as command line parameters. Modi-

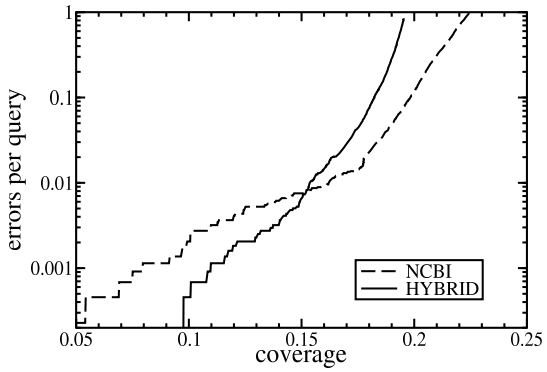

Figure 3. Comparison of the NCBI and the Hybrid version of PSI-BLAST. The curves show the trade-off between the errors per query and the coverage for NCBI and Hybrid PSI-BLAST on a "gold standard" database. For small coverages Hybrid PSI-BLAST is slightly superior while for high coverages the NCBI PSI-BLAST performs better.

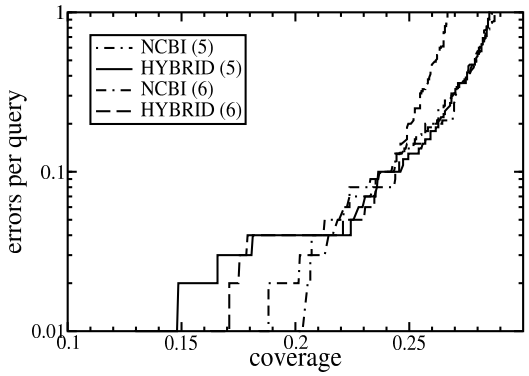

Figure 4. Comparison of the NCBI and the Hybrid version of PSI-BLAST on a large database PDB40NRtrim. The curves show the trade-off between the errors per query and the coverage for NCBI and Hybrid PSI-BLAST for those sequence pairs the homology of which is known from structural considerations. For small coverages Hybrid PSI-BLAST is slightly inferior while for high coverages the two algorithms perform nearly identically.

fying the gap costs affects the gap distribution in the model built in the first iteration, and therefore exposes potential differences in the gap bias of the two algorithms over the following iterations. The measurements result in a family of curves shown in Figure 2. Comparing these curves shows mainly that the Hybrid version of PSI-BLAST is relatively robust with respect to the gap costs. However, among all these rather similar curves, the default value of $11/1$ for NCBI PSI-BLAST seems also optimal for the Hybrid version, suggesting no differences in gap bias.

The final result of this direct comparison between the Hybrid and NCBI version of PSI-BLAST using the gap cost $11/1$ is shown in Figure 3. The curves show that the sensitivity versus selectivity tradeoff of the two versions is quite comparable. Hybrid PSI-BLAST is slightly better than the NCBI PSI-BLAST up to a level of coverage of about 15%, and then incurs slightly more errors than the NCBI PSI-BLAST. The two curves are qualitatively similar, which suggests that their differences reflect the untuned performance of Hybrid PSI-BLAST.

In the second sensitivity assessment we aimed at comparing the two algorithms in a more realistic setting. Instead of searching the very small gold standard database alone, we augmented the gold standard database with the non-redundant protein database from NCBI. Searching this much larger database allows better sequence models to be built and is closer to a typical applications of a tool like PSI-BLAST. The sequences from the gold standard database were marked so that they could be identified from the program output. Sequences in the nonredundant database longer than 10 kilobases were trimmed to 10 kilobases because the protein sequence formatting program 'formatdb' associated with PSI-BLAST 2.0 cannot handle such long sequences. The newly combined dataset was called PDB40NRtrim. Since an exhaustive test using all sequences from the gold standard database as queries would be too time consuming, we randomly selected 100 queries from the gold standard database and searched against PDB40NRtrim. The list of queries is available upon request from the authors. By selecting very high E-value thresholds for output of sequences we ensured that enough of the sequences from the gold standard databases were included in the hit lists. In typical applications of PSI-BLAST, the number of iterations is restricted to a relatively small number since a failure to converge fast is usually a sign of the model being infested by foreign sequences in which case more iterations actually worsen the quality of the model. In order to get an idea of the influence of the maximal number of parameters we chose a limit of 5 and 6 for both algorithms and compared the results. For all other parameters their respective default values were used.

Running the two programs on the PDB40NRtrim dataset took a total of about 64 hours for the NCBI PSI-BLAST, and 54 hours for HYBRID PSI-BLAST. We ran each program on four nodes of a Linux cluster of 1GHz Pentium III, 1GB RAM machines by manually partitioning the list of query sequences equally among the nodes. The use of a cluster reduced the duration of the experiments to a more manage-

able 13-17 hours. Coincidentally, this approach points to an easy way of parallelizing the PSI-BLAST code; along these lines, in a seperate experiment we have written a simple MPI wrapper that enables us to run NCBI tools in parallel on a cluster. As expected, the computational effort of the two algorithms applied to this database of realistic size is comparable, with the HYBRID algorithm taking roughly 25% longer than the original PSI-BLAST. This result confirms that the large difference in computational effort between the HYBRID algorithm and the original PSI-BLAST seen in the short database test is attributable to the startup phase of the HYBRID algorithm.

The sensistivity was assessed by the same curves for the tradeoff between errors per query and coverage as before. In calculating the errors per query and the coverage all hits from the non redundant database were ignored since their homologies are not known. Only hits from the gold standard database were evaluated. The results for the two algorithms for the different limits on the number of iterations are shown in Figure 4. We find, that the HYBRID algorithm seems to depend stronger on the limit on the number of iterations than the original PSI-BLAST. In general the HYBRID algorithm is inferior at small coverages. Note, however, that the region of coverages and errors per query in which the HYBRID algorithm was found to be superior on the smaller database cannot be probed in this test due to the smaller number of queries which limit the errors per query to a minimum of 0.01. At higher coverages the sensitivity of the two algorithms becomes nearly indistinguishable at least if the number of iterations is limited to five.

6 Conclusion

In this study we have established that the hybrid alignment algorithm can be successfully used within PSI-BLAST with only modest changes to the original code. In studying how to best match the algorithm to the PSI-BLAST code, we have resolved the question of sequence length correction. Through direct comparison, and in optimizing one parameter for the hybrid algorithm within the whole framework of PSI-BLAST, namely the gap costs, we found that the Hybrid version of PSI-BLAST and the original version of PSI-BLAST are very similar in their performance.

By incorporating the hybrid alignment directly into the existing PSI-BLAST code, we demonstrate the suitability of the algorithm to the iterative search method where its features can be the most advantageous. We do so in a way that effectively leverages all the efforts that went into the development of the current PSI-BLAST and the tools that build upon it.

This finding will provide a basis for future exploitation of features of the hybrid algorithm that the Smith-Waterman algorithm does not provide. Most notably, it opens the possibility of including position-specific gap costs in PSI-BLAST. Position-specific gap costs represent different propensities for alignment gaps in different regions of the sequences. The propensity for gaps, i.e., for the insertion or deletion of amino acids, is higher in loop regions of a protein family than in its core regions. Thus, it is expected that taking this information into account would greatly improve the sensitivity of PSI-BLAST. Currently, PSI-BLAST is prevented from taking advantage of this additional information due to a fundamental limitation of the underlying theory of the alignment score statistics for Smith-Waterman alignment. This limitation is overcome by the hybrid algorithm, and the results presented here lay the ground for developing a hybrid based version of PSI-BLAST with position-specific gap costs.

References

[1] S. F. Altschul, R. Bundschuh, R. Olsen, and T. Hwa. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research*, 29(2):351–361, January 2001.

[2] S. F. Altschul and W. Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[4] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer. The pfam contribution to the annual nar database issure. *Nucleic Acids Research*, 28(1):263–266, January 2000.

[5] S. E. Brenner, C. Chothia, and T. J. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationship. *Proc. Natl. Acad. Sci. USA*, 95(11):6073–6078, May 1998.

[6] R. Bundschuh. An analytic approach to significance assessment in local sequence alignment with gaps. In S. Istrail *et al*, editor, *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 86–95, New York, New York, 2000. ACM press.

[7] R. Bundschuh. Rapid significance estimation in local sequence alignment with gaps. In S. Istrail *et al*, editor, *Proceedings of the fifth annual international conference on Computational molecular biology*, pages 77–85, New York, New York, 2001. ACM press.

[8] R. Bundschuh. Rapid significance estimation in local sequence alignment with gaps. *J. Comp. Biol.*, 9(2):243–260, April 2001.

[9] J. M. Chandonia, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. Astral compendium enhancements. *Nucleic Acids Research*, 30(1):260–263, January 2002.

[10] J. F. Collins, A. F. W. Coulson, and A. Lyall. The significance of protein sequence similarities. *CABIOS*, 4(1):67–71, March 1988.

[11] L. Conte Lo, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Research*, 28(1):257–259, January 2000.

[12] A. Dembo, S. Karlin, and O. Zeitouni. Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, 22(4):2022–2039, October 1994.

[13] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89(22):10915–10919, November 1992.

[14] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.*, 87(6):2264–2268, March 1990.

[15] S. Karlin and A. Dembo. Limit distributions of the maximal segmental score among markov-dependent partial sums. *Adv. Appl. Prob.*, 24(1):113–140, 1992.

[16] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, November 1998.

[17] P. Koehl and M. Levitt. The astral compendium for sequence and structure analysis. *Nucleic Acids Research*, 28(1):254–256, January 2000.

[18] S. Mercier. *Statistiques des scores pour l'analyse et la comparaison de sequences*. PhD thesis, Université Rouen, 1999.

[19] R. Mott. Maximum likelihood estimation of the statistical distribution of smith-waterman local sequence similarity scores. *Bull. Math. Biol.*, 54(1):59–75, January 1992.

[20] R. Mott. Accurate formula for p-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, 300(3):649–659, July 2000.

[21] R. Mott and R. Tribe. Approximate statistics of gapped alignments. *J. Comp. Biol.*, 6(1):91–112, April 1999.

[22] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, April 1995.

[23] R. Olsen, R. Bundschuh, and T. Hwa. Rapid assessment of extremal statistics for gapped local alignment. In T. Lengauer *et al*, editor, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 211–222, Menlo Park, California, 1999. AAAI press.

[24] W. R. Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. *Genomics*, 11(3):635–650, November 1991.

[25] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85(8):2444–2448, April 1988.

[26] A. B. Robinson and L. R. Robinson. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. USA*, 88(20):8880–8884, October 1991.

[27] A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005, July 2001.

[28] A. A. Schäffer, Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind, and S. F. Altschul. Impala: matching a protein sequence against a collection of psi-blast-constructed position-specific score matrices. *Bioinformatics*, 15(12):1000–1011, December 1999.

[29] D. Siegmund and B. Yakir. Approximate p-values for sequence alignments. *Ann. Stat.*, 28(3):657–680, June 2000.

[30] S. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489, 1981.

[31] T. F. Smith, M. S. Waterman, and C. Burks. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research*, 13(2):645–656, January 1985.

[32] J. L. Spouge. Finite-size corrections to poisson approximations of rare events in renewal processes. *J. Appl. Probab.*, 38(2):554–569, 2001.

[33] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. USA*, 91(11):4625–4628, May 1994.

[34] M. S. Waterman and M. Vingron. Sequence comparison significance and poisson approximation. *Stat. Sci.*, 9(3):367–381, August 1994.

[35] Y. K. Yu, R. Bundscuh, and T. Hwa. Hybrid alignment: High performance with universal statistics. *Bioinformatics*, 18(6):864–872, June 2002.

[36] Y. K. Yu and T. Hwa. Statistical significance of probabilistic sequence alignment and related local hidden markov models. *J. Comp. Biol.*, 8(3):249–282, June 2001.