Suboptimal Alignments Improve the Detection of Weak Homologs in Sequence Database Searches

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

Ralf Bundschuh
Department of Physics
The Ohio State University
191 West Woodruff Avenue
Columbus, OH 43210-1117 U.S.A.
bundschuh@mps.ohio-state.edu

Abstract

PSI-BLAST remains one of the popular tools for searching remote homologs in sequence databases. We recently demonstrated that hybrid alignment can function as the alignment core for PSI-BLAST without loss of sensitivity. Here, we start to exploit the benefits of hybrid alignment. We show that incorporating information about the suboptimal alignments, otherwise ignored in PSI-BLAST, already improves the sensitivity of our enhanced version of PSI-BLAST. More interestingly, we find a set of sequences on which our tool disagrees with the classification given by SCOP. Careful examination points to a possible misclassification in SCOP. Cross-referencing with two other methods of protein structure classification, CATH and DALI, supports this view, indicating that the enriched information from suboptimal alignments is valuable for detecting more weakly homologous sequences.

KEYWORD: sequence alignment, hybrid algorithm, PSI-BLAST, suboptimal alignment, forward-backward algorithm

1 Introduction

Much research has been devoted to understanding evolutionary relationships among biological entities such as genes, regulatory sequences and proteins. These studies provide many opportunities to further our understanding of the structural and functional properties of these biological entities. Sequence alignment remains perhaps the most fundamental approach involved in revealing such biological relationships. Although decades of research and development in sequence alignment analysis has made sequence alignment a well-established technique, the expanding number of genomes encompassing a wider evolutionary history and the need to detect weaker and weaker sequence homology require continuous improvement in the sensitivity of alignment algorithms.

The Smith-Waterman algorithm probably has been the most widely used algorithm in sequence database searches. It always finds the optimal local alignment between two sequences and has been implemented in the database search tool, SSEARCH [27, 23]. Popular bioinformatics tools, such as BLAST [2] and FASTA [24], incorporate heuristic versions of the Smith-Waterman algorithm in order to make large sequence database searches more practical. Recently, Yu and Hwa proposed a variation of the Smith-Waterman algorithm known as hybrid alignment [31]. In contrast to the Smith-Waterman algorithm, hybrid alignment is backed by a theory of statistics that allows one to quickly and reliably assign E-values for arbitrary scoring systems, including position-specific scoring systems. This feature is particularly relevant in iterative approaches, such as PSI-BLAST [2] or SAM [16], which dynamically adapt their scoring systems. These iterative search tools outperform the non-iterative searchs performed by BLAST and FASTA [21].

The hybrid algorithm has the same computational complexity as the Smith-Waterman algorithm and has been combined with the heuristic approaches of BLAST, rendering the hybrid algorithm computationally efficient [30]. It has also been tested as the alignment core of a well established iterative search framework, namely PSI-BLAST [17]. The performance of Hybrid PSI-BLAST compares well with the original PSI-BLAST in detecting sequence homologs.

In this paper, we extend our study of the hybrid algorithm in PSI-BLAST. Motivated by the work of Levitt et al. [12], which demonstrated that the use of suboptimal alignments

improves both the structure homology modeling as well as the sequence-structure alignment, we specifically explore the information contained in suboptimal alignments, which PSI-BLAST ignores. Since not only the optimal alignment, but also the suboptimal alignments, contribute to the final alignment score reported by the hybrid algorithm, it is a natural extension for the hybrid algorithm to search the suboptimal alignment space when building a new model from the sequences found in previous iterations.

We use a database derived from SCOP in order to evaluate our new approach. We compare our new approach to NCBI PSI-BLAST and the previous hybrid PSI-BLAST (Hybrid PSI-BLAST). In examining the false positives reported by the extended hybrid PSI-BLAST (extHybrid PSI-BLAST), we find that many of them are classified as true positives according to two other popular protein classification systems, namely CATH [22] and DALI [13]. We discuss the discrepancy in the classification of these ambiguous sequences and its effect on the evaluation of sequence database search tools.

The remainder of this paper is organized as follows. Section 2 presents some background on sequence alignment statistics. Section 3 provides a general review of the forward-backward algorithm, which we use to exploit suboptimal alignments in refining sequence models. In Section 4, we describe the implementation of the forward-backward algorithm in Hybrid PSI-BLAST and discuss the choice of the pseudocount constant. Section 5 evaluates the extended Hybrid PSI-BLAST approach, and gives a performance comparison between the hybrid and the original version of PSI-BLAST. The differences in the classification of some ambiguous sequences from three protein structure classification methods, namely SCOP, CATH and DALI, is also discussed therein. Finally, Section 6 concludes the paper and proposes directions for future research.

2 Review of sequence alignment statistics

Pairwise sequence alignment algorithms assign a score to the alignment of each pair of sequences. Generally, a larger score implies a closer biological relationship. Iterative sequence alignment tools, such as PSI-BLAST or SAM, build on these pairwise sequence alignment algorithms. In each iteration the pairwise sequence alignment algorithm is used to search a large sequence database leading to a list of hits ordered by their scores. From the high scoring alignments, a multiple alignment is created. That, in turn, determines the scoring system of the next iteration. The crucial step between iterations is deciding which of the hits to keep as putative members of the family (and thus include in the multiple alignment) and which of the hits to reject as irrelevant. A reliable quantitative criterion for this decision is a cutoff in the E- or p-value.

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. Therefore, 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, 11] that the expected number of gapless local alignments of two sequences of length M and N with a score larger than Σ, i.e., the E-value, follows, in the limit of infinitely long sequences, the universal form

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

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 only local alignments are considered. However, the two parameters λ and K do depend on the scoring parameters. The Karlin-Altschul theory [14, 15, 11] 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, to detect weak sequence homologies, gaps must be allowed [23]. According to many numerical studies [10, 18, 29, 1, 20], in the presence of gaps the E-values still follow the universal form Eq. (1). However, the numerical values of the two parameters λ and K are not known.

There are various approaches to solving this dilemma. For large gap costs, approximate analytical formulas exist for λ [19, 26]. For a small sub-class of scoring systems, an analytical formula for λ that is valid for all gap costs [6] has been derived. The current version of PSI-BLAST uses a heuristic method to estimate λ for different scoring matrices but at fixed gap cost [2, 25]. In addition, there are numerical approaches [7, 8] that 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 [31]. 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 score assigned to a sequence pair is from the summation over all the possible alignments instead of from the most probable one alignment as in the Smith-Waterman algorithm. Nevertheless the E-values are still calculated according to Eq. (1) 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 [30]. 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 [30] for a large range of scoring systems with position-specific gap costs taken from the PFAM [3] database. The inability to calculate E-values for position-specific gap costs is precisely the reason why PSI-BLAST does not adopt a this feature, 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 enable us to utilize the suboptimal alignments and open up the possibility to the future incorporation of more sensitive position-specific gap costs.

3 Forward-backward Algorithm

The forward-backward algorithm, which is also called Baum-Welch algorithm, was originally proposed by Baum [4] in studying natural language processing. It can be used to estimate the parameters of a Hidden Markov Model (HMM). With the introduction of HMMs to the sequence alignment research, the forward-backward algorithm has been widely used in many bioinformatics tools. As the name suggests, it consists of two parts: the forward algorithm and the backward algorithm. Both of them can be implemented by dynamic programming just as the Smith-Waterman algorithm and the Needleman-Wunsch algorithm. In fact, the forward algorithm is rather close to the Needleman-Wunsch algorithm in that the forward algorithm just replaces the "max" operation in the Needleman-Wunsch algorithm with the "sum" operation. It thus calculates the sum of the probabilities of all alignments as opposed to just the probability of the optimal alignment. This change will be trivial if the probability of the optimal alignment is dominant, however, it is expected to capture more information if there are many suboptimal alignments with comparable probabilities to the probability of the optimal alignment. The latter case is very common for aligning two remote homologs. The forward algorithm computes the sum of the probabilities of all alignments ending in node j of the HMM and letter i of the sequence for all pairs (i, j). The backward algorithm uses a similar dynamic programming scheme as the forward algorithm to calculate the sum of the probabilities of all alignments that start at node j of the HMM and letter i of the sequence. By coupling the forward and backward algorithm, the posterior probability of the occurence of any amino acid at every model position can be estimated, a process called posterior decoding.

Here is a brief summary of the forward-backward algorithm in the context of hybrid alignment. Let Ω be the model of length m, and $X = (x_1 x_2 \cdots x_n)$ be the sequence to be aligned with the model. On the alignment lattice, the forward algorithm can be used to calculate the total probability $f^S_{i,j}$ of all alignment paths reaching any state S at any position (i, j), where $S \in \{M(match), I(insert), D(delete)\}, i = 1 \ldots n, j = 1 \ldots m$ via the following recursions:

$$\begin{aligned} f^M_{i,j} &= 1 + \eta_{j-1}\omega_{j-1}(x_{i-1})(f^M_{i-1,j-1} \\ &\quad + \mu^{I_1}_{j-1} f^I_{i-1,j-1} + \mu^{D_1}_{j-1} f^D_{i-1,j-1}) \\ f^I_{i,j} &= \mu^{I_2}_j f^M_{i-1,j} + \nu^I_j f^I_{i-1,j} \\ f^D_{i,j} &= \mu^{D_2}_{j-1} f^M_{i,j-1} + \nu^D_{j-1} f^D_{i,j-1} \end{aligned} \tag{2}$$

The boundary conditions are: $f^M_{0,j} = f^M_{i,0} = 1$ and $f^I_{0,j} = f^D_{i,0} = 0$, where $i = 0 \ldots n, j = 0 \ldots m$. The parameter $\omega_j(x_i)$ is the ratio of the emission probability q_{j,x_i} of amino acid x_i over its background probability p_{x_i} at model position j, i.e., $\omega_j(x_i) = q_{j,x_i}/p_{x_i}$. The transition probabilities $\eta_j, \mu^{I_1}_j, \mu^{I_2}_j, \mu^{D_1}_j, \mu^{D_2}_j, \nu^I_j$ and ν^D_j satisfy the following constraint:

$$\begin{aligned} \eta_j + \mu^{I_2}_j + \mu^{D_2}_j &= 1 \\ \eta_j \mu^{I_1}_j + \nu^I_j &= 1 \\ \eta_j \mu^{D_1}_j + \nu^D_j &= 1 \end{aligned} \tag{3}$$

Using the forward algorithm we can find the position (s_E, m_E) at which $Z_{i,j}$ is maximum across the entire lattice, where $Z_{i,j} = f^M_{i,j} + \mu^{I_1}_j f^I_{i,j} + \mu^{D_1}_j f^D_{i,j}$

Once we chose (s_E, m_E) to be the end of the alignment, the backward algorithm can be used to calculate the total probability $b^S_{i,j}$ of all alignment paths starting from state S at position (i, j) and ending at the point (s_E, m_E), where $1 \le i \le s_E, 1 \le j \le m_E$ via the recursions:

$$\begin{aligned} b^M_{i-1,j-1} &= \eta_{j-1}\omega_{j-1}(x_{i-1}) b^M_{i,j} + \mu^{I_2}_{j-1} b^I_{i,j-1} \\ &\quad + \mu^{D_2}_{j-1} b^D_{i-1,j} \\ b^I_{i-1,j-1} &= \mu^{I_1}_{j-1}\eta_{j-1}\omega_{j-1}(x_{i-1}) b^M_{i,j} + \nu^I_{j-1} b^I_{i,j-1} \\ b^D_{i-1,j-1} &= \mu^{D_1}_{j-1}\eta_{j-1}\omega_{j-1}(x_{i-1}) b^M_{i,j} + \nu^D_{j-1} b^I_{i-1,j} \end{aligned} \tag{4}$$

The boundary conditions are: $b^M_{s_E,m_E} = 1$, $b^I_{s_E,m_E} = \mu^{I_1}_{m_E}$, $b^D_{s_E,m_E} = \mu^{D_1}_{m_E}$ and $b^S_{s_E+1,j} = b^S_{i,m_E+1} = 0$, where $i = 0 \ldots s_E, j = 0 \ldots m_E$.

The probability of seeing amino acid A at model position j in sequence X can then be obtained as follows:

$$Pr(A, j, X|\Omega) = \frac{\sum_{x_i=A} f^M_{i,j} * b^M_{i,j}}{Z_{s_E,m_E}} \tag{5}$$

This probability is then summed over all the training sequences. After proper normalization over the 20 amino acids, the new emission probability of the occurrence of any amino acid at position j is computed.

4 Extending Hybrid PSI-BLAST

As part of a previous work [17], a position-specific version of the hybrid algorithm was implemented in version 2.0 of NCBI PSI-BLAST by replacing the Smith-Waterman algorithm with the hybrid algorithm for assigning scores to the sequence hits and assessing the statistical significance of the scores in PSI-BLAST. Hybrid PSI-BLAST has essentially the same user interface with only a small number of new options added for hybrid alignment and retains many of the features of the original NCBI PSI-BLAST.

Hybrid PSI-BLAST calculates the alignment score of every subject sequence using the hybrid algorithm, i.e., the forward algorithm described above. However, once the sequences are selected to be included in the model for the new round, the Smith-Waterman algorithm is executed in order to construct the optimal pairwise alignment between each such sequence and the query. The weighted number of times an amino acid A occurs in column j of the multiple alignment induced by these pairwise alignments yields the normalized frequency $q_{j,A}$, which is later used to determine the substitution score at position j for amino acid A. This process is identical for Hybrid PSI-BLAST and regular PSI-BLAST.

It is fairly natural to extend Hybrid PSI-BLAST to integrate the information about the suboptimal alignments in model building, since Hybrid PSI-BLAST already uses such information in the calculation of the alignment score assigned to each sequence pair. Thus, instead of simply counting how often an amino acid A occurs in column j of the multiple alignment, we will use the posterior probabilities $P(A, j, X|\Omega)$ to determine the contribution of amino acid A in sequence X to column j of the multiple alignment. If the optimal alignment between sequence X and model Ω is dominant, its probability will be the leading term in $P(j, X|\Omega)$. In this case, $P(A, j, X|\Omega)$ has a single peak that is close to 1 for the amino acid A appearing in the optimal alignment and 0 for the other 19 amino acids. Thus we simply "count" the dominant amino acid in this position in the same way as PSI-BLAST and Hybrid PSI-BLAST. Otherwise, the distribution of $P(A, j, X|\Omega)$ is flatter. A single sequence will contribute more than one amino acid to the multiple alignment at column j albeit with a reduced weight reflecting the uncertainty of the alignment at this position.

Since we already implemented the forward algorithm to calculated the alignment scores, we just replaced the routines responsible for calculating the alignment weight matrix by an implementation of the backward algorithm according to Eq. (4) and the decoding of the posterior probabilities through Eq. (5) for the amino acids at each model position for every sequence hit that is selected to build the matrix. The new emission probabilities are then obtained by $q_{j,A} = \frac{\sum_X Pr(A,j,X|\Omega)}{\sum_{X,a} Pr(a,j,X|\Omega)}$

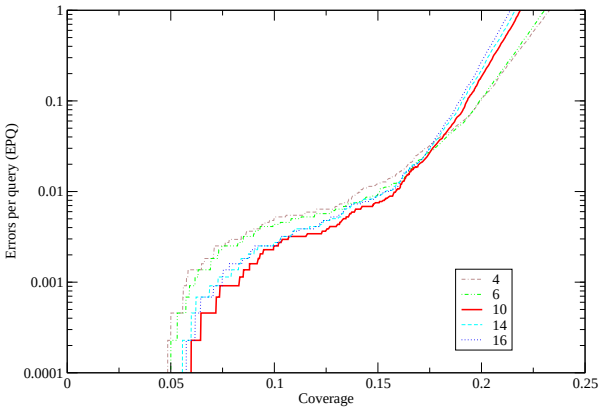


Figure 1. Performance comparison for extHybrid PSI-BLAST with different pseudocount constants.

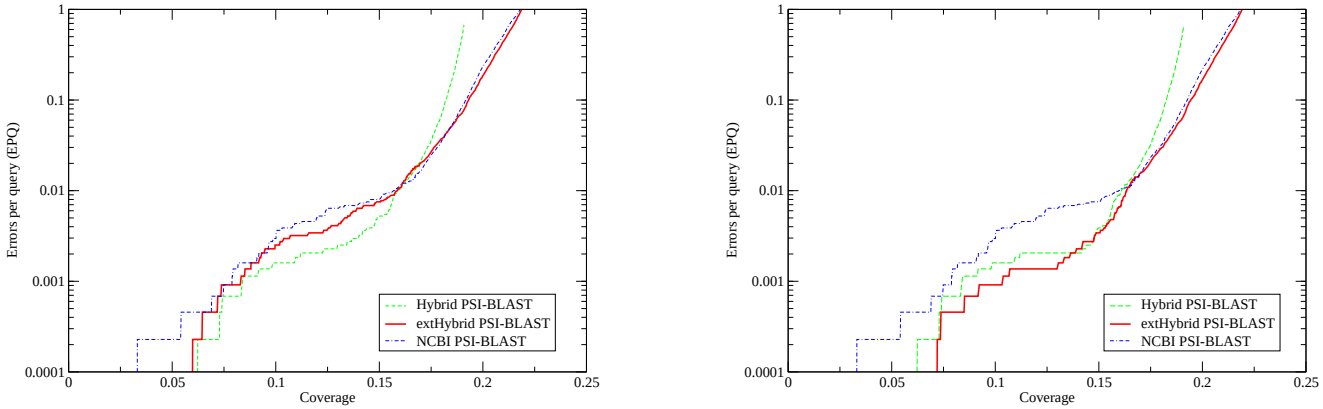
Usually we are limited by the number of the aligned sequences – especially during the first iteration – some variation in the occurrence of certain amino acids may well be missed. Thus it is likely that a matrix derived directly from the $q_{j,A}$ may overfit the data. To address this problem, we recruited the pseudocount approach used in NCBI PSI-BLAST to alleviate the effect of the incompleteness of the sequence sample. In NCBI PSI-BLAST, the actual target frequency of amino acid A at position j is computed as follows:

$$q\prime_{j,A} = \frac{\alpha q_{j,A} + \beta g_{j,A}}{\alpha + \beta} \quad (6)$$

where $g_{j,A} = \sum_b q_{j,b} e^{\lambda_\mu s_{A,b}}$. α and β are pseudocount constants, controlling the weights of the prior knowledge about the target frequencies and the actual observed values from the sample. Here, $s_{A,b}$ is the alignment score for amino acid pair A and b in some nonposition-specific matrix, such as BLOSUM62. λ_μ is the parameter λ in Eq. (1) for the gapless alignment. It is given as the solution of the equation $\sum_{A,b} p_A p_b e^{\lambda_\mu s_{A,b}} = 1$ where p_A and p_b are the background probabilities of amino acid A and b, respectively.

5 Hybrid versus NCBI comparison

In order to accurately evaluate the performance of various sequence alignment algorithms or tools, it will be ideal



(a) straight counting

(b) conservative counting

Figure 2. Performance comparison using different counting methods

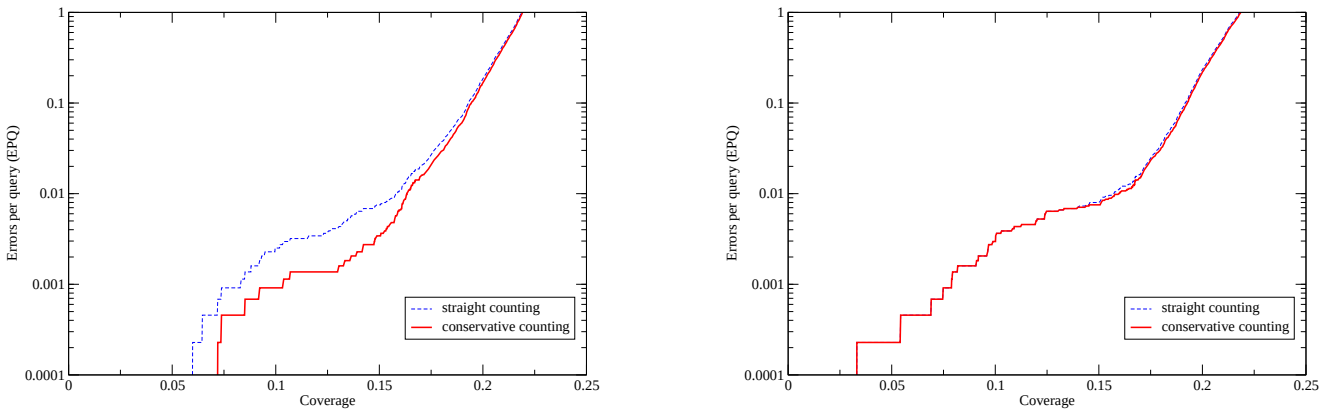
to have a database in which all the relationships among sequences are known. But this ideal database never exists because it is impossible to trace back millions and billions of years to verify the true evolutionary history of the proteins or other biological sequences. Thus, we can only approximate the true relationships by inferring homology from structural and functional clues found so far. SCOP is among the early attempts to create such a database and its classification largely relies on the judgment of human experts. In many ways, SCOP has been considered as the standard of protein structure classification, and thus it has been used in many studies for evaluation of sequence alignment algorithms [5].

In assessing the performance of extHybrid PSI-BLAST, we followed this method. The astral compendium (ASTRAL SCOP 1.59, http://astral.berkeley.edu/) provides a database that contains only sequences with less than 40% pairwise sequence identit [9]. We use all sequences in the database as queries to carry out an all-vs-all comparison. However, one sequence was excluded since we suspect that its true relationship may not be correctly reflected in the SCOP classification. This sequence, namely the representative of the superfamily c.11.1. was consistently misclassified by all versions (Hybrid and NCBI) of the algorithms for nearly all parameter choices (as it turns out that the newest release of ASTRAL indeed changes its classification to c.10.3, which is even a different fold from the old assignment). There are then 4382 sequences with 88171 pairs of true homologos in the remaining database. After pooling together the reported hits and ranking them by E-value, we counted the number of true homologs and the number of non-homologs below various cutoff E-values. For each E-value cutoff, the Coverage is computed by dividing the number of true homologs by the total number of true relationships in the database, which indicates how good the program can recover homologies in the database. On the other hand, the Errors Per Query (EPQ), which is the quotient of the number of non-homologs and the total number of queries, tells us how likely the program would make a mistake. The plot of EPQ versus Coverage as a parametric function of the cutoff E-value demonstrates the tradeoff between the sensitivity and selectivity of the program.

One consideration we had about the extHybrid PSI-BLAST is the choice of the pseudocount constant. Since NCBI PSI-BLAST has undergone years of optimization, it is not clear whether the pseudocount constant used by NCBI is also optimal for HYBRID. Specially we expected that we might have to add less pseudocount since the contribution of the suboptimal alignments already provide counts for amino acids that are absent in the optimal aligments. Thus, we tried different values of the pseudocount constant β (since only the ratio β/α enters in Eq. (6), it is enough to vary one of these constants). We found that the default value for NCBI PSI-BLAST is also optimal for extHybrid PSI-BLAST, as shown in Figure 1.

Having determined the pseudocount for calculating the position-specific alignment weight matrix, we went ahead to evaluate our new approach by comparing the extHybrid PSI-BLAST with Hybrid PSI-BLAST and NCBI PSI-BLAST. In order to assess the changes in sensitivity resulting from the incorporation of the information of the suboptimal alignments, we used two different evaluation schemes. One of them, which we will refer to as straight counting from here on, is the one used previously [5, 8] where two sequences are considered homologs if they are members of the same superfamily of SCOP and non-homologs if not.

However, a concern about the accuracy of the SCOP classification has been recently raised by some researchers. More specifically, the concern is that some of the proteins that are classified in different superfamilies in SCOP might



(a) extHybrid PSI-BLAST

(b) NCBI PSI-BLAST

Figure 3. Performance comparison between two counting methods for NCBI PSI-BLAST and extHybrid PSI-BLAST, respectively.

actually still be homologs. This concern becomes critical as sequence comparison algorithms become sensitive enough to actually find such very weak homologies. Under the straight counting scheme, an algorithm that finds such a weak homolog would be penalized since it would be classified as an error according to the straight counting. To circumvent this difficulty, some suggested that any hit that is in a different superfamily but in the same fold as the query should be ignored instead of being counted as a non-homolog, because we do not have sufficient information to make the judgement of whether such a hit corresponds to a true homolog or not. Here, we refer to this latter counting method as conservative counting.

We applied both counting approaches to evaluate the performance of extHybrid PSI-BLAST and NCBI PSI-BLAST. The result in Figure 2(a) exhibits that extHybrid PSI-BLAST is comparably or slightly more sensitive than NCBI PSI-BLAST in detecting homologous sequences in the database when straight counting is applied. When coverage is low, extHybrid PSI-BLAST makes more errors than Hybrid PSI-BLAST but still fewer than NCBI PSI-BLAST. However, extHybrid improves greatly over Hybrid PSI-BLAST when coverage is high. The picture becomes quite different when conservative counting is applied. Figure 2(b) clearly shows that extHybrid PSI-BLAST outperforms both NCBI PSI-BLAST and Hybrid PSI-BLAST in this case.

Figure 3 highlights the difference between straight and conservative counting for NCBI PSI-BLAST and extHybrid PSI-BLAST, respectively. It can be easily seen that the counting method makes hardly any difference for NCBI PSI-BLAST. This implies that all false hits of NCBI PSI-BLAST are real non-homologs. On the contrary, for extHybrid PSI-BLAST there is a noticeable disagreement between the results for straight counting and for conservative counting. Thus, we suspect that the performance of extHybrid PSI-BLAST is already beyond what the SCOP classification with straight counting can measure, i.e., extHybrid PSI-BLAST can find some homologs that are so remotely related that they are classified into different superfamilies in SCOP.

To further augment this point, we collected hits with E-value less than 0.01 for extHybrid PSI-BLAST that are false hits under the straight counting scheme but undecided according to conservative counting. We examined their relationships by referring to another popular protein classification database CATH [22], which combines manual and automated processes to organize the proteins. There are 39 ambiguous sequence pairs found by extHybrid PSI-BLAST that are classified as non-homolog pairs in SCOP. In CATH, almost all of these proteins pairs (35 out of 39) are classified as homologs. The exceptions are 3 sequence pairs that are not classified at all in CATH and one sequence pair (d1jtdb_ vs. d1k3ia3, the one with the largest E-value of 0.009) that is classified as non-homologous. These 39 sequence pairs involve 31 sequences, listed in Table 1.

To further investigate the possible homologies uncovered by extHybrid PSI-BLAST, we also submitted the coordinate files of the 35 pairs that are considered homologs in CATH to DaliLite, which is the structure comparison and database search engine for another protein classification system DALI. The result shows that except for two very short sequences ($\sim$ 50 amino acids in length) the Z-score for each pair is above 10. As for the root-mean-square deviation (RMSD) of α-carbon atoms, 22 pairs share structures within 3Å RMSD, 11 pairs within 4Å RMSD and 2 pairs

Table 1. Ambiguous sequences

SCOP NAME	SCOP ID	CATH ID	SCOP NAME	SCOP ID	CATH ID
d1hg3a_	c.1.1.1	3.20.20.90	d1a4ya_	c.10.1.1	3.80.10.10
d1thfd_	c.1.2.1	3.20.20.90	d1yrga_	c.10.1.2	3.80.10.10
d1rpxa_	c.1.2.2	3.20.20.90	d1fqva2	c.10.1.3	3.80.10.10
d1dbta_	c.1.2.3	3.20.20.90	d1h6ta2	c.10.2.1	3.80.10.10
d2tpsa_	c.1.3.1	3.20.20.90	d1h6ua2	c.10.2.1	3.80.10.10
d2dora_	c.1.4.1	3.20.20.90	*d1jl5a_	c.10.2.6	3.80.10.10
d1ep3a_	c.1.4.1	3.20.20.90	d1dcea3	c.10.2.2	3.80.10.10
d1d3ga_	c.1.4.1	3.20.20.90	d1tfi__	g.41.3.1	2.20.25.10
d1gox__	c.1.4.1	3.20.20.90	d1qyp__	g.41.9.1	2.20.25.10
d1ltda1	c.1.4.1	3.20.20.90	d1i50i2	g.41.9.1	2.20.25.10
d1h7wa2	c.1.4.1	3.20.20.90	d1en2a2	g.3.1.1	N/A
d1ea0a2	c.1.4.1	N/A	d1fjna_	g.3.7.3	N/A
d1zfja1	c.1.5.1	3.20.20.90	d1hf2a1	b.80.3.1	N/A
d1ak5_1	c.1.5.1	3.20.20.90	d1ea0a1	b.80.4.1	N/A
d1jr1a1	c.1.5.1	3.20.20.90	d1jtdb_	b.69.5.2	2.130.10.30
			d1k3ia3	b.69.1.1	2.130.10.80

within 6Å RMSD. These findings provide a strong evidence for these sequence pairs to be classified as homologs.

6 Conclusion

In this paper, we have extended our previous work on the hybrid algorithm by incorporating information about suboptimal alignments in the refining process of the sequence models in PSI-BLAST. This is achieved by implementing the forward-backward algorithm in place of the routines for building the alignment matrix in HYBRID PSI-BLAST. We also experimented with various values for the pseudocount constant and found that the default value for NCBI PSI-BLAST is optimal for extHybrid PSI-BLAST as well.

It turns out that extHybrid PSI-BLAST and NCBI PSI-BLAST are very close in their performance, with extHybrid PSI-BLAST being slightly more sensitive than NCBI PSI-BLAST if the regular evaluation method using the SCOP classification is employed. However, if we take some caution in the classification of the SCOP database and use a conservative counting method in evaluating the performance, we find that extHybrid PSI-BLAST outperforms NCBI PSI-BLAST, primarily due to the differences in the false positives that are identified by these two alignment programs, i.e., the "errors" NCBI PSI-BLAST makes are mostly true errors, where as many of the "errors" extHybrid PSI-BLAST makes are potential homologs. For the collection of those "errors" made by extHybrid PSI-BLAST with reported E-value less than 0.01, we found that most of them are classified as homologs in CATH and DALI. We conclude that extHybrid PSI-BLAST reaches a sensitivity that exceeds what can be reliably measured by the SCOP classification.

Here we only investigate one feature provided by hybrid alignment. There are many attributes available to the hybrid algorithm but lacking in the Smith-Waterman algorithm which have not been explored. The most prominent one is the ability to handle position-specific gap costs, which takes into account the different indel (insertion/deletion) propensities along the protein sequences. It is expected that it is more likely to have an indel in loop regions of a protein family than in its core regions, which has a more stable structure. This information is thought to be valuable in improving the sensitivity of the alignment algorithm, but is not utilized in PSI-BLAST because of a fundamental limitation of the underlying theory of the alignment score statistics for the Smith-Waterman alignment. Based on the work described in this paper, it is possible for us to extend the forward-backward algorithm for calculating the gap probabilities in a position-specific fashion and to examine the potential effect on searching remotely related sequences.

7 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0317335. The authors would like to thank Stephen Altschul and Nicholas Chia for helpful discussion.

References

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

[2] 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.

[3] 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.

[4] L. E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3(627):1–8, 1972.

[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 and et. al., editors, *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 and et. al., editors, *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] 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.

[12] M. Gerstein and M. Levitt. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology*, pages 59–67, Menlo Park, California, 1996. AAAI press.

[13] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233(1):123–138, September 1993.

[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] Y. Li, M. Lauria, and R. Bundschuh. Using hybrid alignment for iterative sequence database searches. *Concurrency and Computation: Practice and Experience*, 9(16):841–853, August 2004.

[18] 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.

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

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

[21] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, 284(4):1201–1210, December 1998.

[22] F. Pearl, A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, T. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, and C. Orengo. The cath domain structure database and related resources gene3d and dhs provide comprehensive domain family information for genome analysis. *Nucleic Acids Research*, 33:247–251, January 2005.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

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

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

[31] 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.