5966–5973 *Nucleic Acids Research, 2006, Vol. 34, No. 20* *Published online 26 October 2006*
doi:10.1093/nar/gkl731

Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches

Yi-Kuo Yu, E. Michael Gertz, Richa Agarwala, Alejandro A. Schäffer and Stephen F. Altschul*

National Center for Biotechnology Information, National Library of Medicine, NIH, DHHS, Bethesda, MD 20894, USA

Received July 27, 2006; Revised September 15, 2006; Accepted September 21, 2006

ABSTRACT

Protein sequence database search programs may be evaluated both for their retrieval accuracy—the ability to separate meaningful from chance similarities—and for the accuracy of their statistical assessments of reported alignments. However, methods for improving statistical accuracy can degrade retrieval accuracy by discarding compositional evidence of sequence relatedness. This evidence may be preserved by combining essentially independent measures of alignment and compositional similarity into a unified measure of sequence similarity. A version of the BLAST protein database search program, modified to employ this new measure, outperforms the baseline program in both retrieval and statistical accuracy on ASTRAL, a SCOP-based test set.

INTRODUCTION

Computer programs that search protein or DNA databases often are evaluated by their ability to distinguish true biological relationships from chance similarities. Such an evaluation requires a test query set and database for which the true relationships are known. Given a query sequence, a database search program returns alignments in some specific order, often ranked using an objective measure such as an alignment score, *E*-value or *P*-value. One way to evaluate search accuracy is with ROC (receiver operating characteristic) analysis (1). To produce a ROC curve, the number of true positives is plotted against the number of false positives returned as one descends the retrieval list. The ROC_n score, the normalized area under this curve up to the first n false positives, has become a popular measure of search accuracy. ROC_n scores may be calculated for individual queries or, if objective scores render distinct database searches comparable, the search results from many different queries may be pooled to produce a single combined ROC curve and score (2).

Although ROC_n scores capture one aspect of search method quality, they ignore an important issue. In practice, most database searches return results for which the true and false positives are not known. Even a perfect ordering, with all the true results preceding all the false ones, may be of limited use if a user has no effective method for drawing a line between the two classes. Most protein database search programs now provide, in addition to an ordered list, *E*- or *P*-values with which to assess whether any given result can be explained by chance. Many applications rely critically upon these values for further analyses. For example, PSI-BLAST (3) and related iterative protein profile search programs use an *E*-value threshold to automatically include alignments in further rounds of analysis; a single false positive below this threshold can corrupt all subsequent results (4). Although unreliable *E*-values may be countered by setting a very stringent threshold, this can squander efforts that have gone into improving retrieval accuracy. Accordingly, as important a consideration for a program's utility as the degree to which it separates true from chance similarities may be the accuracy with which it calculates *E*-values (5).

A central question in calculating alignment *E*-values is defining the random distribution to which they refer. The simplest approach is to calculate *E*-values with reference to a random protein model, based on standard amino acid frequencies; this is the baseline behavior of the BLAST programs (3,6). The most immediate problem arises from 'low-complexity' segments—protein regions with extremely restricted amino acid usage—which depart drastically from the random model. It is usually not of interest to align such segments, and they may be filtered out of consideration using a program such as SEG (7). However, even after low-complexity segments have been removed, many proteins have distinctly biased amino acid compositions. Such biases are typical of some protein families, but particular organisms also have AT- or CG-biased genomes, leading by means of the genetic code to characteristically biased proteomes (8,9). Accordingly, many authors have proposed calculating the significance of an alignment of two proteins by considering their amino acid compositions within some model

*To whom correspondence should be addressed. Tel: +301 435 7803; Fax: +301 480 2288; Email: altschul@ncbi.nlm.nih.gov

for generating random sequences (2,10,11). If one does not account for compositional bias, the reported *E*-values or *P*-values may be orders of magnitude too low. As we show below, an implementation of this basic idea within BLAST greatly improves the accuracy of its statistical evaluations.

One might expect that adopting a more accurate calculation of *E*-values would yield improved retrieval accuracy as well. However, retrieval accuracy actually decreases, even when the search results from many queries are pooled (2). A possible explanation is that similar amino acid compositional biases for two proteins constitutes, in itself, some evidence of protein relationship. The challenge, then, is to produce a mathematically justified way of taking compositional similarity into account when assessing sequence relationships.

The approach we take here is to consider two distinct measures of sequence similarity: the traditional measure of alignment similarity and a new measure of compositional bias similarity. We investigate empirically the distribution of compositional similarity among unrelated proteins. This allows us to define an associated compositional *P*-value distinct from an alignment *P*-value. We find that, on average, related proteins have a greater compositional similarity than unrelated proteins. Furthermore, we find that for unrelated sequence pairs, alignment and compositional *P*-values are effectively independent. Therefore, we can combine these *P*-values into a single, unified *P*-value (12–14), which can serve as a new measure for assessing sequence similarity. We show that this measure recaptures the previously forfeited retrieval accuracy, while at the same time yielding accurate statistics. By also employing the previously described compositional score adjustment (15–17), we generate a program that substantially outperforms the baseline BLAST program on an ASTRAL test set (18,19) both in retrieval accuracy and in the accuracy of its reported *E*-values.

THEORY

Variants of BLAST

We will compare five variants of the gapped protein-query, protein-database BLAST program (3), which we distinguish in this paper with different prefixes: B-BLAST, S-BLAST, SU-BLAST, C-BLAST and CU-BLAST (Table 1). The baseline program B-BLAST is modified in a few minor ways from the default protein–protein BLAST program available on the web site of the National Center for Biotechnology Information (NCBI) (20). As noted below, some of these changes are for testing purposes only, in order to minimize the number of confounding factors when comparing B-BLAST to the other BLAST variants, while others may be retained in future web versions of BLAST.

Table 1. Summary of the five variants of BLAST considered in this study

Program	Scoring adjustment	Unified *P*-values
B-BLAST	None	No
S-BLAST	Compositional Scaling	No
C-BLAST	Compositional Adjustment	No
SU-BLAST	Compositional Scaling	Yes
CU-BLAST	Compositional Adjustment	Yes

Compositional Scaling refers to the method in (2) and Compositional Adjustment refers to the method in (15).

The program S-BLAST scales the scores of a standard matrix, for each alignment reported, based upon the compositions of the two sequences compared. This approach, which improves statistical evaluations, was introduced by Schäffer *et al.* (2). The program SU-BLAST, which is new to this paper, combines the alignment similarity of S-BLAST with compositional similarity, described below, to produce a unified measure of sequence similarity.

The program C-BLAST conditionally adjusts the scores of a standard matrix, for each alignment reported, based upon the compositions of the two sequences compared. The approach is described in Altschul *et al.* (15), and is based on methods from Yu *et al.* (17) and Yu and Altschul (16). The program CU-BLAST, which is new to this paper, combines the alignment similarity of C-BLAST with compositional similarity, described below, to produce a unified measure of sequence similarity.

All five variants use the program SEG (7) to filter database sequences for low-complexity regions. SEG replaces certain amino acids with the character 'X', which is also used to signify an unknown amino acid. Past default versions of BLAST have assigned a fixed negative score to the aligned pair (α, X), where α is a standard amino acid. Here, all five variants assign to (α, X) a weighted average of the scores for α aligned to the twenty standard amino acids. The new way to score aligned letter pairs involving X may be retained in future default versions of BLAST. Here, the calculated composition of any sequence that is filtered using SEG ignores those amino acids that are replaced with 'X'.

All five variants use the optional Smith-Waterman algorithm (21) to generate all local alignments actually reported. Also, in this final alignment step, all five variants use five more bits of precision for their substitution scores than are used by the standard BLOSUM-62 matrix (22). In both respects, B-BLAST differs from past default versions of BLAST.

Except for its extra precision in the final step, B-BLAST uses the standard BLOSUM-62 matrix in conjunction with scores of $-11-k$ for gaps of length k. For its *E*-value calculations, it employs gapped statistical parameters that are estimated (23) for a set of standard amino acid frequencies (24). S-BLAST uses 'composition-based statistics' (2) to scale the BLOSUM-62 substitution scores for any pair of sequences, while leaving the gap scores fixed. The program SU-BLAST, introduced here, combines a measure of compositional bias similarity with S-BLAST's measure of alignment similarity to produce a unified measure of sequence similarity. C-BLAST and CU-BLAST replace the scaled BLOSUM-62 scores of S-BLAST and SU-BLAST with the conditionally compositionally adjusted scores described by Altschul *et al.* (15).

There is one additional change we made for testing purposes only. The implementation of composition-based statistics, available for many years on the web-site of the NCBI, places upper and lower bounds of 1.0 and 0.5 on the factor by which a substitution matrix can be scaled. The upper bound is imposed to improve slightly the program's retrieval accuracy and speed (2). The lower bound, rarely invoked, improves the utility of the program's output for certain applications. However, these artificial bounds confound the issues we wish to study here, and so we have

removed the upper bound entirely, and reduced the lower bound to 0.05.

Finally, since the publication of the paper introducing 'conditional compositional adjustment' to derive a new substitution matrix (15), we have found that in addition to the three criteria therein described, a fourth is appropriate for invoking compositional adjustment. Specifically, compositional adjustment should be used for any comparison involving a protein of length at least 50 amino acids and whose two most abundant residues constitute at least 40% of the protein. The implementations of C-BLAST and CU-BLAST studied here employ this additional criterion, but it has no measurable effect on the paper's results.

BLAST heuristics

All variants of BLAST we study involve scaling or adjusting the substitution matrix for each alignment reported. BLAST is a heuristic program, not guaranteed always to find the optimal alignments, and scaling or adjusting the substitution matrix separately for each database sequence would unduly increase execution time. Accordingly, we re-evaluate only those database sequences which pass an initial screen. Specifically, the standard gap scores, BLOSUM-62 matrix, and statistical parameters for standard amino acid frequencies (24) are used to calculate preliminary *E*-values. Any database sequence producing an alignment with a preliminary *E*-value less than or equal to a set threshold, here taken to be 100, is retained for further evaluation. The substitution and gap scores are rescaled or adjusted, and an optimal local alignment is generated using the rigorous Smith-Waterman algorithm.

SU-BLAST and CU-BLAST calculate compositional *P*-values to combine with alignment *P*-values in order to produce unified *P*- and *E*-values for reporting. Because alignments are produced only for database sequences that pass the initial screen, SU-BLAST and CU-BLAST calculate compositional *P*-values for only these sequences.

Statistics

For the comparison of random sequences, an analytic, asymptotic statistical theory has been developed for the distribution of scores of ungapped local alignments (25,26). In brief, for the comparison of two random sequences of lengths m and n, the number of distinct local alignments with score at least S is approximately Poisson distributed, with expected value

$$E = Kmn\,e^{-\lambda S}, \qquad (1)$$

where K and λ are calculable parameters dependent upon the scoring system and amino acid distribution. The Poisson distribution implies that the *maximum* score follows an extreme value distribution (27), with the probability of achieving a score at least S given by

$$P = 1 - e^{-E} = 1 - \exp(-Kmn\,e^{-\lambda S}). \qquad (2)$$

This formula may be inverted, yielding

$$E = -\ln(1 - P). \qquad (3)$$

For ungapped alignments, λ is defined only for scoring systems with negative expected score. It is the unique positive solution to the equation

$$\sum_{1 \leqslant i,\, j \leqslant 20} p_i\, p'_j\, e^{s_{ij}\lambda} = 1\,, \qquad (4)$$

where s_{ij} is the score for aligning amino acids i and j, and p_i and p'_j are the background probabilities, respectively, for amino acid i in the first sequence and amino acid j in the second (25,26). Empirically, for typical scoring regimes, optimal gapped local alignments follow the same type of distribution as do ungapped local alignments (28), although the distribution's parameters λ and K can not be calculated analytically but may be estimated by random simulation (23).

The *E*-value or *P*-value of an alignment score depends upon the lengths of the sequences compared. *E*- or *P*-values may be reported in the context of a pairwise comparison or in the context of a database search. For the database search context, the length n of a single sequence is replaced in formulas 1 and 2 by the aggregate length N, in residues, of all database sequences. By default, the BLAST programs report database *E*-values, but because we will discuss pairwise *E*- and *P*-values below, we will always use a hat symbol to indicate pairwise as opposed to database *E*- or *P*-values. The relationship between the pairwise and database *E*-values for an alignment involving a database sequence of length n (29) is given by the formula

$$E = \frac{N}{n}\hat{E}. \qquad (5)$$

Unified *P*- and *E*-values

As described below, we will combine an alignment pairwise *P*-value $\hat{P}_a$ with a compositional *P*-value P_c to calculate a unified pairwise *P*-value $\hat{P}_u$. Because BLAST reports database *E*-values, it must perform the following calculations to convert an alignment database *E*-value E_a into a unified database *E*-value E_u:

$$\hat{E}_a := \frac{n}{N}E_a \qquad \text{(Equation 5)};$$

$$\hat{P}_a := 1 - e^{-\hat{E}_a} \qquad \text{(Equation 2)};$$

$$\hat{P}_u := f(\hat{P}_a, P_c) \qquad \text{(see below)};$$

$$\hat{E}_u := -\ln(1 - \hat{P}_u) \qquad \text{(Equation 3)};$$

$$E_u := \frac{N}{n}\hat{E}_u \qquad \text{(Equation 5)};$$

E_u is reported, and can be converted into a database *P*-value P_u using Equation 2 if desired.

Program evaluation

We evaluate versions of BLAST both for the reliability of their statistics and for their retrieval accuracy. To study the reliability of the statistics produced, we compare 10 000 shuffled mouse sequences of length at least 150 with shuffled

human RefSeq (20) sequences from Build 35 of the human genome. For each query, we record the lowest database *P*-value returned. We then plot the number of queries for which this *P*-value is $\leqslant x$.

To study retrieval accuracy, we use the 'astral40' data set (18,19), based upon the SCOP structural classification of proteins (30,31), for ROC analysis. Specifically, the 3586 astral40 sequences related to at least one other astral40 sequence are compared to the complete data set, and the results of all searches are pooled by *E*-value. For increasing *E*-value, the number of true positives is plotted against the number of false positives, and a ROC_{5000} score is calculated.

RESULTS

Here, we first study the statistical and retrieval accuracy of the previously described programs B-BLAST and S-BLAST. We then describe a measure of similarly biased amino acid compositions in two proteins and a method for combining this compositional similarity with alignment similarity to create a unified assessment of sequence similarity. We implement this unified measure in SU-BLAST, whose statistical and retrieval accuracies we study. Finally, we replace the compositional scaling employed by S-BLAST and SU-BLAST with conditional compositional matrix adjustment to create C-BLAST and CU-BLAST, and study the performance of these programs. The central idea behind such compositional matrix adjustment is to find target frequencies for aligned amino acid pairs that are consistent with the background frequencies of the sequences being compared, but as close as possible to the target frequencies implicit in a standard substitution matrix. For the comparison of sequences with different background frequencies, the resulting substitution matrix is asymmetric. It is fruitful to employ compositional adjustment only under certain conditions (15), with compositional scaling (2) used when these conditions fail.

B-BLAST and S-BLAST

The proper statistical parameters λ and *K* for the distributions of Equation 1 and 2 depend upon the amino acid compositions of the sequences being compared. Thus the significance of alignments with the same nominal score can vary, dependent upon the context in which the alignments arise. For example, using BLOSUM-62 scores (22), a high-scoring alignment is much more likely to arise by chance from the comparison to two cysteine-rich proteins than from the comparison of two proteins with more typical amino acid compositions. Numerically, this is mediated in Equation 1 primarily by the cysteine-rich compositions implying a smaller value of λ, which discounts the nominal score.

The baseline B-BLAST program evaluates all alignments using gapped statistical parameters λ_g^* and K_g^* estimated (23) for a standard amino acid composition (24). Thus, for alignments involving proteins whose compositions imply a gapped λ_g substantially smaller than λ_g^*, the *E*-values reported may be much smaller than justified.

BLAST estimates the number of alignments that are expected to achieve a given score by chance, i.e. from the comparison of unrelated proteins. Our test of BLAST statistics retains the compositions of real proteins, while scrambling the order of their amino acids. Figure 1 shows that B-BLAST's statistics are far from accurate. For example, in 10 000 randomized B-BLAST database searches, 639 (6.4%) yield best matches with *P*-value $\leqslant 10^{-4}$, when only one would be expected. Furthermore, some queries can yield best matches with extremely inaccurate statistical assessments: 143 queries (1.4%) returned best matches with *P*-value $\leqslant 10^{-10}$. Note also that when the best random match has an extremely low *P*-value or *E*-value, many other matches frequently do as well. For example, a single query whose best match had a B-BLAST *P*-value of 10^{-12} yielded 101 matches with *E*-value $\leqslant 10^{-4}$.

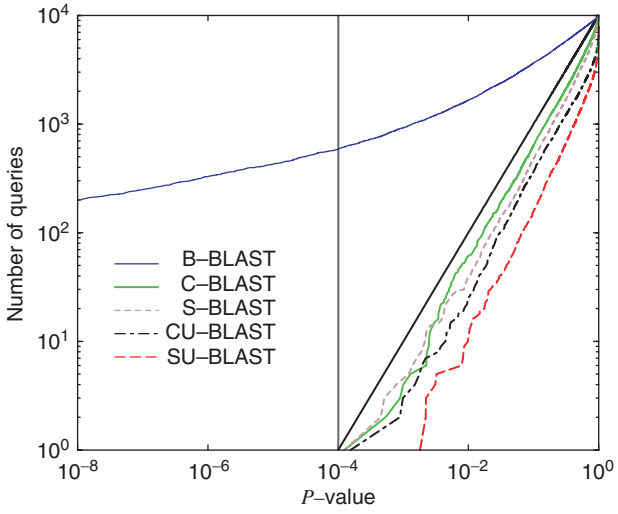

Figure 1. The accuracy of BLAST statistics. 10 000 shuffled mouse sequences were compared to shuffled human RefSeq (20) sequences from Build 35 of the human genome. The number of queries whose best match had a reported *P*-value $\leqslant x$ is plotted against *x*, using a log–log scale. Curves are shown for B-BLAST, S-BLAST, SU-BLAST, C-BLAST and CU-BLAST. The diagonal line indicates the theoretical prediction for all curves. The vertical line at $x = 10^{-4}$ indicates the point at which a single query with equal or better *P*-value is expected.

This problem with BLAST statistics is understood to be due primarily to similarly biased compositions among many protein sequences and is largely mitigated by the 'composition-based statistics' (2) employed by S-BLAST. To estimate rapidly the statistical parameters for gapped alignments, S-BLAST multiplies the standard BLOSUM-62 matrix by a distinct constant for each pair of sequences, so that the scaled matrix has the same ungapped scale parameter λ in the new compositional context that the unscaled matrix has in the standard context. The gap costs remain fixed. When the new scoring system is employed, the optimal local alignment may change. Therefore alignments must be recalculated, as described in the Theory section above, after the substitution matrix has been scaled.

Figure 1 shows that the statistics of S-BLAST are far more accurate that those of B-BLAST, and even slightly conservative. From 10 000 database searches, only six best matches are returned with *P*-value $\leqslant 10^{-3}$, where ten are expected. In some applications it is crucial to exclude false positives reliably, for example when constructing a PSI-BLAST position-specific score matrix for further database searches

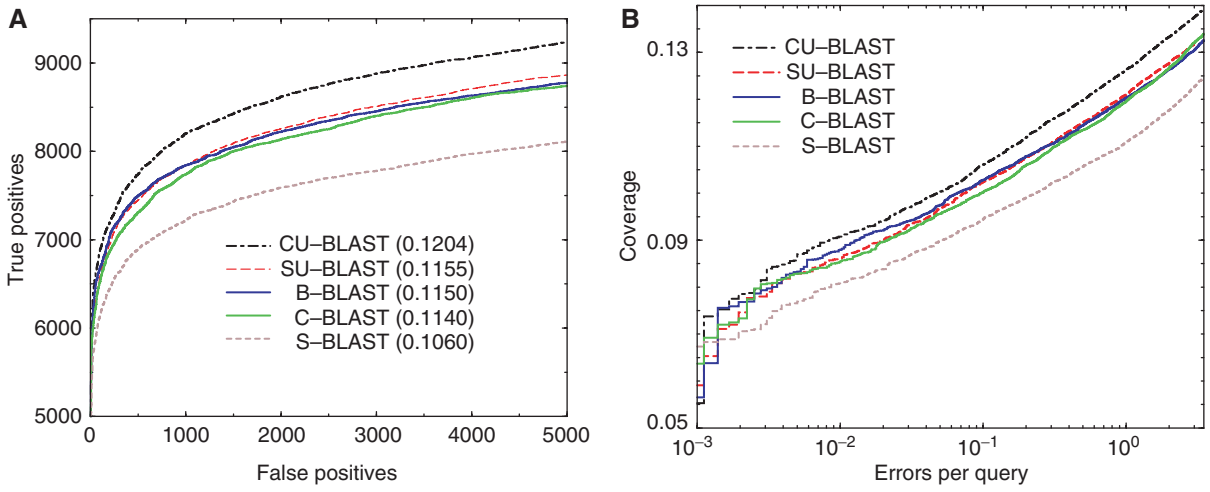

Figure 2. BLAST retrieval accuracy. The 3586 astral40 sequences having at least one other relative in the astral40 data set (18,19) are used as queries in a search of this database. The results are pooled and sorted by *E*-value, and ROC curves are produced by plotting the number of true positives against the number of false positives as one descends the retrieval list. In (**A**), ROC curves are shown for B-BLAST, S-BLAST, SU-BLAST, C-BLAST and CU-BLAST. The ROC_{5000} scores for these programs are also shown, each having a standard error of ± 0.0002 (2). In (**B**), the same ROC curves are shown in a semi-log plot, using the scales of coverage and errors per query (30).

(3). In these instances, S-BLAST is strongly preferred to B-BLAST (2). However, if one ignores the accuracy of reported statistics and pays attention only to the relative abilities of the two programs to separate true from chance similarities, then B-BLAST appears better; Figure 2 shows that its ROC curve lies significantly above S-BLAST's. This seemingly paradoxical result can be understood by recognizing that similarly biased compositions can in themselves constitute evidence for sequence relatedness. This evidence is effectively discarded when one calculates the significance of an alignment given the compositions of the sequences being compared. The problem we now address is whether this evidence can be recaptured in a mathematically justified manner, thereby restoring the retrieval accuracy of B-BLAST while retaining the statistical accuracy of S-BLAST.

Compositional similarity

To study coordinated amino acid biases among related and unrelated proteins, we first require an appropriate measure. The difference between the substitution scores used by B-BLAST and S-BLAST, a difference that caused a sizable erosion in retrieval accuracy, was multiplication by a factor proportional to the ungapped scale parameter λ. We therefore propose to use λ itself, calculated for a fixed reference set of substitution scores, but using the observed amino acid frequencies of two proteins, as a measure of coordinated amino acid bias. An analysis of equation (4) whose solution is λ (26) shows that λ will be low for any pair of proteins with an unusually large number of amino acids having high mutual substitution scores as defined by the reference substitution matrix.

There is no model from which one may derive an accurate distribution of λ for unrelated proteins. Accordingly, we proceed by calculating λ, based on the BLOSUM-62 substitution scores (22), for all pairs of unrelated proteins from the astral40 data set (18,19). The resulting empirical probability density function for λ is shown in Figure 3. We use this

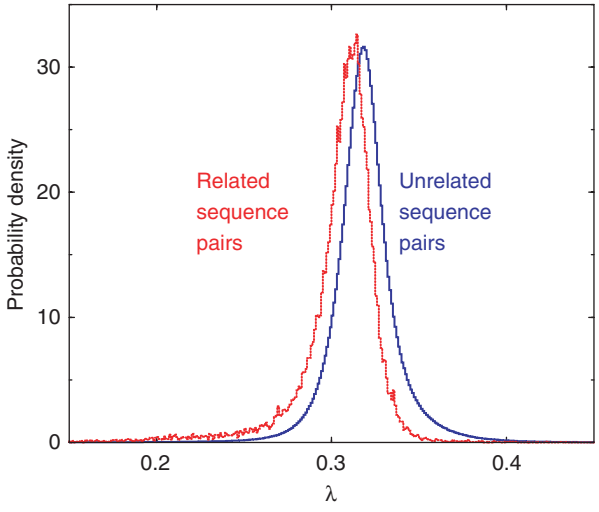

Figure 3. The probability density function for λ. The ungapped scale parameter λ (26) is calculated for the standard BLOSUM-62 amino acid substitution matrix (22) using the observed amino frequencies of two proteins. Empirical probability density functions are shown for all pairs of unrelated proteins from the astral40 data set (18,19), as well as for all pairs of non-identical related proteins.

distribution to assign to any particular value of λ an empirical compositional *P*-value, $P_c(\lambda)$, equal to the area under the density curve to the left of λ. Because for small λ the data supporting our empirical distribution become sparse, we set a lower bound on $P_c(\lambda)$ of 10^{-6}, which is returned whenever λ less than equal to 0.068.

Figure 3 also shows the empirical probability density of λ for pairs of related sequences from the astral40 data set. Our strategy is to glean from the separation between the distributions for related and unrelated sequences evidence of sequence relatedness based on compositional considerations alone.

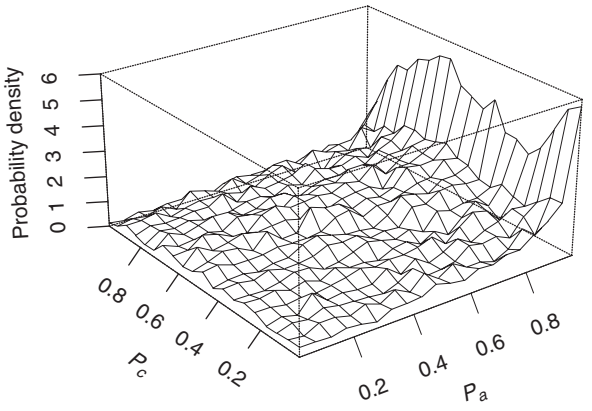

Figure 4. The empirical probability density of alignment and compositional *P*-values from shuffled sequences. 10 000 shuffled mouse query sequences were compared using S-BLAST to shuffled human RefSeq (20) sequences from Build 35 of the human genome. For each query, the alignment *P*-value P_a of the best match to the database was found, and the compositional *P*-value P_c was then calculated for the database sequence involved.

Combining alignment and compositional significance

How may one combine P_c for a pair of sequences with a traditional alignment-score *P*-value, P_a, calculated using composition-corrected substitution scores, to derive a valid unified *P*-value, P_u, based upon both compositional and alignment evidence? One approach is to assume that P_a and P_c are independent random variables with a uniform distribution on the interval (0,1). Define a new random variable equal to the product of P_a and P_c, which has an easily calculated probability density over (0,1), from which a unified *P*-value may be derived (12–14). The resulting P_u is given by

$$P_u = P_a P_c (1 - \ln P_a P_c). \quad (6)$$

For this formula to be valid, P_a and P_c should be accurate and independent. If we calculate P_a using composition-based statistics (2), as described above for S-BLAST, there should be little if any dependence between P_c and P_a. To test whether this is effectively the case, for each best match of a shuffled mouse sequence to a shuffled human RefSeq sequence, we calculated P_a and P_c and plotted their joint empirical probability density in Figure 4. There was close to no evident dependence between the values of P_a and P_c, although P_a was systematically high. For values of $P_a < \sim 0.3$, the probability density shown in Figure 4 is close to uniform, although appreciably < 1.0. This implies that for such P_a, Equation 6 should yield values for P_u that are consistently conservative by a constant multiplicative factor. This is approximately what is observed in Figure 1.

SU-BLAST

We have implemented a program, here called SU-BLAST, that uses both compositional and alignment evidence to assess protein sequence similarity. Before we discuss this program's retrieval and statistical accuracy, two technical details bear comment.

First, because we wish to combine compositional similarity not just with the best hit from a database search, but in principle with the best hits for all sequences, it is appropriate to apply Equation 6 with pairwise *P*-values $\hat{P}_a$ and $\hat{P}_u$ in place of database *P*-values P_a and P_u. Therefore, to calculate a unified database *E*-value E_u from an alignment database *E*-value E_a, we have to perform the five-step calculation described above in the Theory section. Second, as described in the Theory section above, SU-BLAST returns a result only for those sequences whose preliminary alignment *E*-value is lower than a set threshold. This heuristic is unlikely to exclude many alignments with low unified *E*-values because alignment similarity generally carries much more information than compositional similarity.

We tested SU-BLAST for retrieval and statistical accuracy. As shown in Figure 1, SU-BLAST's reported statistics are noticeably more conservative than those of S-BLAST, but are still accurate within an order of magnitude. However, as shown in Figure 2, by using λ to measure and reward similar compositional bias, SU-BLAST recaptures the retrieval accuracy forfeited by S-BLAST. SU-BLAST and B-BLAST have very similar ROC curves and ROC_{5000} scores. For the data set used, B-BLAST is slightly better at false positive rates < 0.3/query, and SU-BLAST is slightly better at false positive rates > 0.3/query. The major difference between SU-BLAST's and B-BLAST's performance is found in the far greater accuracy of SU-BLAST's statistics.

C-BLAST AND CU-BLAST

It has been argued that standard substitution matrices such as BLOSUM-62 are not ideal for comparing sequences with non-standard compositions, and an efficient method has been proposed for adjusting standard matrices for use in arbitrary compositional contexts (15–17).

For protein database searching, Altschul *et al.* (15) have shown that conditional compositional substitution matrix adjustment yields better retrieval accuracy than does the substitution matrix scaling (2) embodied in S-BLAST. Like matrix scaling, compositional adjustment produces alignment statistics conditioned on the compositions of the sequences compared. Therefore, it is appropriate to replace the matrix scaling of S-BLAST and SU-BLAST with conditional compositional adjustment (15) to produce the programs C-BLAST and CU-BLAST. An analysis of the independence of P_a and P_c similar to that shown in Figure 4, but with S-BLAST replaced by C-BLAST, produces results nearly equivalent to those discussed above, and is omitted here.

The reliability of C-BLAST's and CU-BLAST's statistics is evaluated in Figure 1. As can be seen, the replacement of scaled by compositionally adjusted substitution matrices yields a somewhat improved agreement of statistical theory with experiment. Also, as shown in Figure 2, CU-BLAST outperforms both B-BLAST and SU-BLAST in retrieval accuracy. In summary, evaluated from the baseline provided by B-BLAST, the integration of compositionally adjusted substitution matrices with a measure of similar compositional bias yields a program that is substantially improved from the standpoints of both statistical and retrieval accuracy.

DISCUSSION AND CONCLUSION

It has been recognized for some time that a failure to account for biased amino acid compositions can lead to exaggerated

claims of protein alignment statistical significance (2,10,11). It has not been widely recognized, however, that basing alignment statistics upon sequence composition can erode retrieval accuracy. We have argued that such erosion may stem from the fact that similarly biased compositions, in themselves, constitute evidence of protein relatedness. To improve alignment statistics, earlier methods have teased apart alignment and compositional similarity, but have then discarded the latter. We have proposed rejoining these two threads of evidence in a mathematically valid manner. Some studies may involve comparing related proteins that have amino acid compositions known or suspected to be discordant. In such cases, adding to alignment similarity the compositional similarity discussed here may well be counterproductive.

We have derived an empirical distribution of ungapped λ values for unrelated sequence pairs from the astral40 data set (18,19). This set is biased towards globular proteins, and therefore our λ distribution may not be valid for more comprehensive protein sets. Unfortunately, accurate classifications of proteins into related and unrelated classes are not currently available for such larger sets, and the λ distribution may bear refinement as protein relationships become more fully understood. However, proteins with highly biased or repetitive sequences generally are heavily filtered by the SEG program, and are probably not best studied using traditional alignment methods.

Different versions of the command line executable program 'blastpgp' implementing the five BLAST versions described above and compiled for Linux are available at: ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/blast_unified_statistics. Also available in this directory are a table of empirical values for $P_c(\lambda)$, the shuffled human and mouse sequences used in the statistical tests above, as well as the unshuffled sequences from which they are derived.

In this paper we have been concerned primarily with comparing the newly proposed measure of sequence similarity to earlier ones, and have sought to minimize confounding factors introduced by various BLAST heuristics. Accordingly, we have used the Smith-Waterman option in the final alignment phase (2), and we have used a high preliminary *E*-value threshold of 100 for re-evaluating alignments. However, the modest improvements in search sensitivity yielded by these options come at a significant cost in execution time, and different options may be chosen as defaults for NCBI's BLAST web servers. We have no reason to believe that our results would be qualitatively different using either faster BLAST parameter settings, or using the Smith-Waterman algorithm on every database sequence instead of on only those selected by the BLAST heuristics.

The use of compositional similarity, as well as its integration with compositional matrix adjustment, will be available as an option for the protein-query, protein-database BLAST search program on the NCBI web site. The current command line *blastpgp* offers unified *P*-values as part of the composition option (*-t*). Also, the use of compositional similarity may be specified for PSI-BLAST's initial BLAST round of database search, but thereafter PSI-BLAST uses only alignment similarity to assess results.

Compositional sequence relatedness has been used in other ways for protein sequence analysis before. For example, the PHD program for predicting protein secondary structure (32) employs global amino acid composition as one input to a neural network. Also, a somewhat ad hoc approach has been described for adjusting the reported *E*-values of alignments involving low complexity regions by post-processing BLAST outputs (33). Some database search programs, such as FASTA (34), correct for the composition of query sequences by estimating statistical parameters from database searches. This procedure, however, takes no account of the compositions of individual database sequences, and so can at most partially correct for the manner in which compositional biases skew pairwise alignment scores.

It may be possible to improve in several ways on this paper's approach. For example, the statistical corrections for compositional bias employed by S-BLAST and C-BLAST, while quite accurate for random sequences, are compromised to varying degrees by periodicity and non-uniformity within real sequences. SEG (7), applied to database sequences in this paper, removes many low-complexity segments from consideration, but certain periodic patterns remain. These may be dealt with by additional special-purpose filters, e.g. for coiled coils (35–38), or by calculating alignment statistics based on a reversed-sequence model of randomness (10,39).

SCOP-based test sets such as astral40 are widely used for the evaluation and comparison of protein sequence database search methods, but they have potential disadvantages. SCOP is a classification of protein domains, but most comparisons performed in database searches involve complete proteins. As a result, SCOP-based evaluations may tend unduly to favor alignment scoring systems that have more of a global than a local flavor, such as the compositional bias similarity studied here. A test set we have previously employed (2), and which compares queries to full-length yeast sequences, is too small to yield statistically significant results in this study. Certainly the utility of the methods we have discussed is a function of the degree to which the compositional properties of complete proteins reflect those of the proteins' domains. Progress on the automatic parsing of protein sequences into domains should therefore be able to improve both statistical and retrieval accuracy.

Here we have studied one measure, the ungapped λ implied by a fixed substitution matrix, of two sequences' compositional similarity. Many other measures are possible, and we did investigate one, a compositional distance metric recently described by Endres and Schindelin (40). The distance distributions for related and unrelated sequence pairs showed a marked separation, similar to that of Figure 3. However, the improvement in retrieval accuracy yielded by this measure was distinctly inferior to that yielded by λ (data not shown). Nevertheless, it remains possible that theoretical considerations or further experimentation will yield a compositional similarity measure more effective than λ for our purposes.

An alternative approach to measuring global compositional similarity is by log–odds scores, analogous to those for alignment similarity, which make use of information derived from related as well as unrelated sequence pairs (John Spouge, personal communication). One may imagine other methods for taking advantage of the different behaviors of these two sets, and it is likely that a more sensitive measure than those we have so far studied will be found.

The idea of combining alignment similarity with independent measures of sequence relatedness, such as compositional similarity, may be applied fruitfully to database search programs other than BLAST (34). It may also be possible to graft compositional or other similarity measures onto the alignment similarity measures used by protein profile programs such as HMMER (41), PSI-BLAST (3), SAM (10), IMPALA (42) or SALTO (43). To what extent this can improve the statistics or retrieval accuracy of these programs awaits further investigation.

ACKNOWLEDGEMENTS

Thanks to Aleksandr Morgulis for comments on the manuscript and for technical assistance. This research was supported by the Intramural Research Program of the National Library of Medicine of the NIH/DHHS. Funding to pay the Open Access publication charges for this article was provided by the Intramural Research Program of the National Institutes of Health, NLM.

Conflict of interest statement. None declared.

REFERENCES

1. Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.
2. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
3. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.
5. Pearson,W.R. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.
6. 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.
7. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
8. Sueoka,N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci. USA*, **85**, 2653–2657.
9. Wan,H. and Wootton,J.C. (2000) A global compositional complexity measure for biological sequences: AT-rich and CG-rich genomes encode less complex proteins. *Comput. Chem.*, **24**, 71–94.
10. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
11. Mott,R. (2000) Accurate formula for *P*-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, **300**, 649–659.
12. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
13. Elston,R.C. (1991) On Fisher's method of combining *p*-values. *Biom. J.*, **33**, 339–345.
14. Fisher,R.A. (1958) *Statistical Methods for Research Workers*. 13th edn. Hafner, New York, NY, pp. 99–101.
15. Altschul,S.F., Wootton,J.C., Gertz,E.M., Agarwala,R., Morgulis,A., Schäffer,A.A. and Yu,Y.-K. (2005) Protein database searches using compositionally adjusted substitution matrices. *FEBS J.*, **272**, 5101–5109.
16. Yu,Y.-K. and Altschul,S.F. (2005) The construction of amino acid substitution matrices for the comparison of proteins with nonstandard compositions. *Bioinformatics*, **21**, 902–911.
17. Yu,Y.-K., Wootton,J.C. and Altschul,S.F. (2003) The compositional adjustment of amino acid substitution matrices. *Proc. Natl Acad. Sci. USA*, **100**, 15688–15693.
18. Chandonia,J.-M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
19. Green,R.E. and Brenner,S.E. (2002) Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proc. IEEE*, **90**, 1834–1847.
20. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
21. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
22. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
23. Altschul,S.F., Bundschuh,R., Olsen,R. and Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.
24. Robinson,A.B. and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl Acad. Sci. USA*, **88**, 8880–8884.
25. Dembo,A., Karlin,S. and Zeitouni,O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, **22**, 2022–2039.
26. 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.
27. Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York, NY.
28. Smith,T.F., Waterman,M.S. and Burks,C. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.*, **13**, 645–656.
29. Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.
30. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
31. 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.
32. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enyzmol.*, **266**, 525–539.
33. Sharon,I., Birkland,A., Chang,K., El-Yaniv,R. and Yona,G. (2005) Correcting BLAST e-values for low-complexity segments. *J. Comp. Biol.*, **12**, 980–1003.
34. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
35. Berger,B., Wilson,D.B., Wolf,E., Tonchev,T., Milla,M. and Kim,P.S. (1995) Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA*, **92**, 8259–8263.
36. Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enyzmol.*, **266**, 513–525.
37. McDonnell,A.V., Jiang,T., Keating,A.E. and Berger,B. (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, **22**, 356–358.
38. Wolf,E., Kim,P.S. and Berger,B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.*, **6**, 1179–1189.
39. Karplus,K., Karchin,R., Shackelford,G. and Hughey,R. (2005) Calibrating *E*-values for hidden Markov models using reversesequence null models. *Bioinformatics*, **21**, 4107–4115.
40. Endres,D.M. and Schindelin,J.E. (2003) A new metric for probability distributions. *IEEE Trans. Info. Theory*, **49**, 1858–1860.
41. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
42. Schäffer,A.A., Wolf,Y.I., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
43. Kann,M.G., Thiessen,P.A., Panchenko,A.R., Schäffer,A.A., Altschul,S.F. and Bryant,S.H. (2005) A structure-based method for protein sequence alignment. *Bioinformatics*, **21**, 1451–1456.