METHODS: A Companion to Methods in Enzymology
Vol. 3, No. 1, August, pp. 66–70, 1991

Improved Sensitivity of Nucleic Acid Database Searches Using Application-Specific Scoring Matrices

David J. States,[1] Warren Gish, and Stephen F. Altschul

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

Scoring matrices for nucleic acid sequence comparison that are based on models appropriate to the analysis of molecular sequencing errors or biological mutation processes are presented. In mammalian genomes, transition mutations occur significantly more frequently than transversions, and the optimal scoring of sequence alignments based on this substitution model differs from that derived assuming a uniform mutation model. The information from sequence alignments potentially available using an optimal scoring system is compared with that obtained using the BLASTN default scoring. A modified BLAST database search tool allows these, or other explicitly specified scoring matrices, to be utilized in computationally efficient queries of nucleic acid databases with nucleic acid query sequences. Results of searches performed using BLASTN's default score matrix are compared with those using scores based on a mutational model in which transitions are more prevalent than transversions.

© 1991 Academic Press, Inc.

A database search is a query of a complex information resource. Optimal results from a database query depend on asking the most specific question possible. Often when a similarity search is performed, many possible matches are identified, and it becomes necessary to choose among them. For molecular sequence database searches, this is frequently done by assigning similarity scores to sequence alignments and ranking the possible matches on the basis of these scores (1, 2). The scoring system is based on a model of sequence relationship and is implicit in the formulation of the database query. It is therefore important to use scores appropriate to the question of interest.

BLASTN uses a simple scoring system in which matches count +5 and mismatches −4 (2). To achieve computational efficiency, these scores have been incorporated directly into the source code. Although this scoring system is adequate for many tasks, in this paper we describe some situations in which alternative scoring systems are desirable, as well as a method for incorporating such scores into BLAST searches of nucleic acid databases.

Natural mutations do not interconvert the various bases uniformly: transitions are favored over transversions by a factor of approximately 3 (3). Furthermore, the scoring of closely related sequences should differ from that of sequences known to be distantly related. This may be analyzed using a Markov transition model (4), which for protein sequence comparison has formed the basis of the well known PAM matrices (5). Such matrices can be derived for nucleotide sequence comparisons as well (6). We assess their performance relative to that of the BLASTN default scores in seeking homology among noncoding DNA sequences.

In the course of DNA sequencing projects, it is frequently useful to know whether two fragments contain an identical segment of sequence so that they may be assembled into a contig. Although raw DNA sequence data may contain errors, they are far less frequent than the rate for which the default scores in BLASTN are optimized. Alternative scores, designed to search for nearly identical matches, are presented. Such scores may also be useful for evaluating sequence segments to be employed as PCR or oligonucleotide hybridization primers.

LOG-ODDS SCORE MATRICES

Score matrices have been based on log-odds ratios derived from a Markov mutational model (5, 7). Such a model assumes that mutation (substitution) events are random and independent. A matrix M of probabilities for substituting base i by base j after any given amount of evolution can be calculated by successive iteration of a reference mutation matrix:

$$M_n = (M_1)^n. \quad [1]$$

[1] To whom correspondence should be addressed.

1046-2023/91 $3.00
Copyright © 1991 by Academic Press, Inc.
All rights of reproduction in any form reserved.

M_1 is a matrix reflecting 99% sequence conservation and one point accepted mutation (1 PAM) per 100 bases. M_n then represents the substitution probabilities after n PAMs. To model the case for which all base substitutions are equally likely, the diagonal elements of M_1 are all 0.99, while all off-diagonal elements are 0.00333. For a biased mutation model in which a given transition is threefold more likely than a given transversion, the off-diagonal elements of M_1 corresponding to transitions are 0.006 and those for transversions are 0.002.

The n-PAM log-odds score S_{ij} for aligning a given pair of bases is simply the log of the relative chance of that pair occurring as a result of evolution as opposed to that occurring from a random alignment of two bases,

$$S_{ij} = \log\left(\frac{p_i M_{nij}}{p_i p_j}\right), \qquad [2]$$

where p_i is the underlying frequency of base i. The symmetry in the choice of the matrices M_1 described above essentially assumes equal frequencies for the four nucleotides, and S_{ij} can be written as $\log(4\ M_{nij})$.

In order to be statistically significant, an alignment needs to achieve a score of about $\log N$, where N is the size of the search space, i.e., the product of the length of the database (in residues) and the length of the query sequence (7, 8). If the base of the logarithm used in formula [2] is taken to be 2, then scores can be thought of as being expressed in bits, and significance calculations can easily be performed in one's head.

At any given PAM distance, the expected information H per alignment position can be calculated as described by Altschul (7). Assuming that the scores are expressed in bits, then H, also in bits, is given by

$$H = \sum_{i,j} p_i p_j S_{ij} 2^{S_{ij}}. \qquad [3]$$

In order to employ arbitrary nucleotide replacement scores, the program BLASTP (2), modified as discussed below, was used for database searches.

EFFICIENCY OF SCORES BASED ON VARIOUS MUTATIONAL MODELS

Table 1 shows the log-odds scores derived using the uniform substitution model for various PAM distances. These scores are expressed in bits, but for computational purposes the scores may be multiplied by any positive number. At 30 PAMs (about 75% sequence conservation when back mutations are considered) the magnitudes of the match and mismatch scores are nearly identical, and at 47 PAMs the ratio is approximately 5 to 4. Scaled by a constant factor, these are the scores incorporated into BLASTN. The average information per alignment position is also shown in Table 1. For example at 47 PAMs, the BLASTN default, about 0.5 bit of information per aligned base is obtained. By comparison, if it is known a priori that the sequences are highly similar, say differing by a 1% sequencing error rate, then 1.9 bits per base can be obtained by using an optimal scoring matrix. Conversely, if the sequences are highly diverged, then much less information per aligned base is available; at 100 PAMs divergence, only 0.13 bit remain.

Of course, the PAM distance corresponding to an alignment cannot be known before the alignment is found, but the information $H(D)$ available at various PAM distances D (shown in Table 1) is achieved only if the appropriate scores are used. Since one does not want to use hundreds of different scoring systems, an important question is over what range of actual PAM distances a given set of scores is nearly optimal. Using a set of scores optimized for PAM distance E, it is simple to calculate the average score achieved when segments actually separated by PAM distance D are aligned. We call the ratio of this score to $H(D)$, which will always be less than or equal to 1, the *efficiency* of the PAM scores E at PAM distance D. The efficiency curves for various PAM ma-

TABLE 1

PAM Substitution Scores Based on the Uniform Mutation Model

PAM distance	Percentage conserved	Match score (bits)	Mismatch (score (bits)	Match/ mismatch score ratio	Average information per position (bits)
1	99.0	1.99	−6.24	0.32	1.90
2	98.0	1.97	−5.25	0.38	1.83
5	95.2	1.93	−3.95	0.49	1.64
10	90.6	1.86	−3.00	0.62	1.40
15	86.4	1.79	−2.46	0.73	1.21
20	82.4	1.72	−2.09	0.82	1.05
25	78.7	1.66	−1.82	0.91	0.92
30	75.3	1.59	−1.60	0.99	0.80
35	72.0	1.53	−1.42	1.07	0.70
40	69.0	1.46	−1.27	1.15	0.62
45	66.2	1.40	−1.15	1.22	0.54
50	63.5	1.34	−1.04	1.29	0.47
55	61.0	1.29	−0.94	1.36	0.42
60	58.7	1.23	−0.86	1.43	0.37
65	56.5	1.18	−0.79	1.50	0.32
70	54.5	1.12	−0.72	1.56	0.28
75	52.6	1.07	−0.66	1.62	0.25
80	50.8	1.02	−0.61	1.68	0.22
85	49.1	0.97	−0.56	1.74	0.19
90	47.6	0.93	−0.52	1.80	0.17
95	46.1	0.88	−0.48	1.85	0.15
100	44.8	0.84	−0.44	1.90	0.13
105	43.5	0.80	−0.41	1.96	0.12
110	42.3	0.76	−0.38	2.01	0.10
115	41.2	0.72	−0.35	2.05	0.09
120	40.1	0.68	−0.33	2.10	0.08
125	39.2	0.65	−0.30	2.14	0.07

trices (E equal to 10, 30, 50, 70, and 90) are graphed in Fig. 1. Note that the curve for PAM scores E always achieves a maximum of 1.0 at PAM distance $D = E$. The degree of efficiency considered necessary will depend to some extent on the size of the search conducted (7). For example, if 30 bits of information are required for statistical significance, to get within 2 bits of the optimal achievable score will require an efficiency of close to 94%, while if only 16 bits are required one can get within 2 bits of the optimal with an efficiency of 89%. Given a desired degree of efficiency and a desired range of actual PAM distances, one may calculate how many different PAM matrices need be employed and which ones should be used. For example, as can be seen from Fig. 1, using both PAM-30 and PAM-70 scores gives over 92% efficiency for actual PAM distances anywhere from 10 to 90. The 90% efficiency range of the PAM-47 scores used as a default by BLASTN is from actual PAM distances 20 to 68.

Table 2 shows a series of scoring matrices derived from a biased mutational model in which each transition is three times more likely than each transversion. Mismatches here are scored differently, depending upon which type of mutation they represent. Interestingly, at greater than 87 PAMs, transitions score positively and are therefore conservative substitutions. The efficiency curves for scores based on this model are similar to those based on the uniform mutational model, but somewhat flatter. Thus, a given set of scores will be useful over a somewhat greater range of actual PAM distances.

If the mutations in the DNA sequences being compared are biased in the manner of this model, then the scores of Table 2 distinguish true relationships from random noise more efficiently than those of Table 1. Calculation shows that for PAM distances D from 0 to 100, using scores for the correct PAM distance but based on the uniform as opposed to biased mutational model yields an efficiency of about $(100 - D/5)\%$. This can amount to a substantial loss of information for alignments of sequences that have diverged by more than 40 PAMs. As an example, consider the alignment shown in Fig. 2, one of several found in a BLASTP search of GenBank (Release 68.0) with the PAM-50 scores of Table 2, using the 5′ upstream region of the human p53 gene [Accession No. J04238 (9)] as a query. (All scores were multiplied by 100 so that the program could work with integers.) The alignment involves the 5′ flanking region of the rat tumor antigen p53 gene [Accession No. M26863 (10)] and represents a true homology. Containing 41 matches, 13 transitions, and 7 transversions, the alignment has a total score of 39.75 bits; in the context of the search performed (query sequence length 532; database length 65, 868, 799) this translates to a P value of 0.013. If the PAM-50 scores from Table 1 are used in place of those from Table 2, the alignment's score drops to 34.88 bits and its first three positions are trimmed. This constitutes an information loss of over 12%, and raises the P value of the alignment

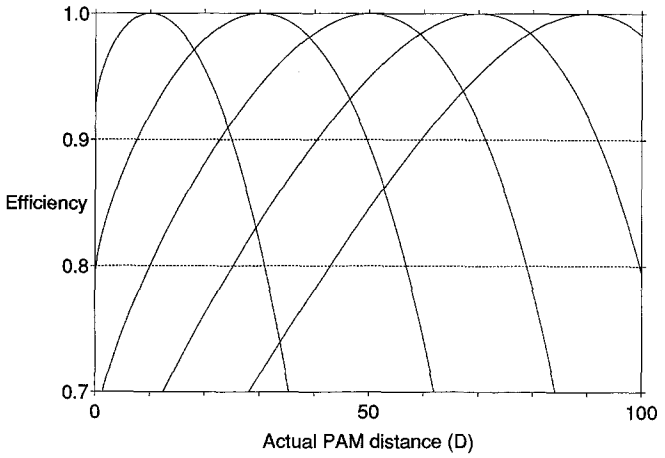

FIG. 1. Efficiency of various uniform mutational model PAM matrices (E equal to 10, 30, 50, 70, and 90), as a function of the actual PAM distance D of the sequences being compared. Scores based on PAM distance E have their maximum efficiency (1.0) when the actual PAM distance $D = E$.

TABLE 2

PAM Substitution Scores Based on the Biased Mutation Model

PAM distance	Percentage conserved	Match score (bits)	Transition score (bits)	Transversion score (bits)	Average information per position (bits)
5	95.2	1.93	−3.13	−4.67	1.65
10	90.7	1.86	−2.19	−3.70	1.42
15	86.5	1.79	−1.67	−3.14	1.24
20	82.6	1.72	−1.32	−2.76	1.09
25	79.0	1.66	−1.06	−2.46	0.96
30	75.6	1.60	−0.86	−2.23	0.85
35	72.4	1.54	−0.70	−2.03	0.76
40	69.5	1.48	−0.57	−1.87	0.67
45	66.8	1.42	−0.47	−1.73	0.60
50	64.2	1.36	−0.37	−1.60	0.54
55	61.8	1.31	−0.30	−1.49	0.48
60	59.6	1.25	−0.23	−1.39	0.43
65	57.5	1.20	−0.17	−1.30	0.39
70	55.6	1.15	−0.12	−1.22	0.35
75	53.8	1.10	−0.08	−1.15	0.32
80	52.1	1.06	−0.04	−1.08	0.29
85	50.5	1.01	−0.01	−1.02	0.26
90	49.0	0.97	0.02	−0.96	0.23
95	47.6	0.93	0.04	−0.91	0.21
100	46.3	0.89	0.06	−0.86	0.19
105	45.1	0.85	0.08	−0.82	0.17
110	44.0	0.81	0.10	−0.77	0.16
115	42.9	0.78	0.11	−0.73	0.14
120	41.9	0.74	0.12	−0.70	0.13
125	41.0	0.71	0.13	−0.66	0.12
130	40.1	0.68	0.14	−0.63	0.11
135	39.2	0.65	0.15	−0.60	0.10
140	38.5	0.62	0.16	−0.57	0.09
145	37.7	0.59	0.16	−0.54	0.08
150	37.1	0.57	0.16	−0.52	0.08

to 0.31. The BLASTN default PAM-47 scores perform comparably.

THE USE OF BLASTP FOR NUCLEIC ACID SEARCHES

Nucleic acid database searches with application-specific scores are easily implemented using the BLASTP program (2). The program can read a user-defined substitution score matrix and has command line options to tailor the search algorithm to specific tasks. Minimal source code modifications are needed to adapt BLASTP for use in nucleic acid searching. The array ("fq") containing the expected database residue frequencies must be modified to reflect nucleotide rather than amino acid frequencies; in the present work, uniform frequencies for A, C, G, and T were used. Also, the maximum sequence length ("QUERYLEN_MAX") needs to be increased to accommodate the longer sequences encountered in nucleic acid databases.

Command line options should also be utilized to optimize BLASTP performance for nucleic acid searches. Increasing "W", the word size for the neighborhood table, to 6 will dramatically improve the speed of the search. Setting "T", the score threshold for including a word in the neighborhood table, to a large positive value will yield a table that contains only matches. Finally, with the score matrix specified in hundreths of bits, increasing "X", the cutoff for extending word hits, to 1000 will reduce the probability of prematurely truncating an aligned segment to less than 0.1%. By default, X is adjusted heuristically.

There is a computational cost to the increased flexibility achieved using BLASTP rather than BLASTN. Because BLASTN is restricted to a four-character alphabet, employs hard-coded scores, and uses a long word size (12), the actual search phase is faster. The most important of these factors is the long word size in BLASTN, but this causes a significant loss of sensitivity for moderately diverged sequences. For example, the alignment shown in Fig. 2 is missed altogether by BLASTN because it lacks a run of 12 identities needed to generate a hit in the BLASTN neighborhood table.

To achieve maximum specificity, the sense and antisense strands are searched separately by BLASTP, allowing orientation information such as that available from a cDNA or directional promoter element to be utilized. In contrast, BLASTN automatically searches both strands of the query.

SCORES FOR DETECTING SEQUENCE OVERLAPS

In the course of sequencing projects it is frequently useful to know whether a new segment of sequence overlaps an existing sequence significantly enough to form the basis of a contig. This question may be addressed by creating a database out of the existing sequence segments and comparing the novel segment to this database. In this case, one is interested only in alignments that differ by sequencing errors. Sequencing substitution errors are probably uniformly distributed and may occur at a rate as high as 2 to 5 per 100 bases of raw sequence data. PAM-5 scores are over 96% efficient for PAM distances 0 to 12, and so would be a reasonable choice for this sort of data, while PAM-47 scores are only about 70% efficient in the lower part of this range. (An additional virtue of PAM-5 scores is that they can be written essentially as +1 for a match and −2 for a mismatch; to convert to bits one need only multiply by 1.92.) If the query sequence is 300 bases long and the existing data set contains 100 fragments each 300 bases long, then $\log_2(100^* 300^* 300)$ or about 24 bits of information will be needed to achieve significance; an extra 6 or 7 bits are needed if 99.9% confidence is required. At a distance of 2 PAMs, about 1.83 bits of information per aligned base are optimally available so that using an efficient matrix overlaps of average length 13 to 17 bases will suffice. In contrast, using the BLASTN default PAM-47 scores, a significant signal would need to contain on average 18 to 23 aligned bases. Since insertion/deletion mutations are present in addition to substitutions, the reduced size of an uninterrupted segment needed by an efficient matrix to achieve significance may substantially improve the ability to recognize contigs.

PROTEIN CODING REGIONS

It has been observed that when a DNA sequence codes for protein, it is generally more fruitful to search a protein sequence database with the translated sequence than to

```
Human p53:   74 GATCCAGCTGAGAGCAAACGCAAAAGCTTTCTTCCTTCCACCCTTCATATTTGACACAATG 134
                  G  CCA  T A A   A CG AAAAGCTT  TTC TTCC C CTT  TA TTGACACA TG
Rat p53:     84 GGCCCACTTAAAATAGATCGTAAAAGCTTAATTCTTTCCGCTCTTTTTACTTGACACAGTG 144
```

FIG. 2. An alignment of the 5′ upstream regions for human and rat p53 genes. Identities are echoed on the central line. Using the biased mutation model PAM-50 scores of Table 2, this alignment has a score of 39.75 bits, which renders it significant (P value = 0.013) in the context of a search of GenBank (Release 68.0). Using the uniform mutation model PAM-50 scores of Table 1, the alignment's score drops to 34.88 bits, which is not statistically significant (P value = 0.31).

search a nucleic acid sequence database directly. The development above allows us to provide a rough quantitative analysis of why this is the case.

Consider two proteins that have diverged by D protein PAMs. This involves D nonsynonymous point mutations at the DNA level. Li *et al.* (3) have shown that, broadly speaking, there tend to be over 1.5 synonymous point mutations for every nonsynonymous point mutation. Therefore, because there are 3 nucleotides per codon, each amino acid PAM translates into roughly $(1 + 1.5)/3 \approx 0.8$ nucleic acid PAMs. Table 3 shows the average information available in protein sequence alignments at various PAM distances (7). Each of these distances has been translated into a corresponding nucleic acid PAM distance, and Table 3 then shows the average information available per codon, based on the biased mutational model discussed above. From these data one may assess whether the protein or nucleic acid alphabet (used naively, without reference to the genetic code) carries more information. It will be seen that for alignments of sequences that have diverged by fewer than 50 protein PAMs the nucleic acid alphabet is more informative, while for more distant relationships the protein alphabet is superior.

For searches of current protein databases, the most typical distance of alignments that are just distinguishable from chance is approximately 120 protein PAMs (7). Table 3 shows that at this distance, about 37% of the information available through an amino acid substitution matrix is lost using a nucleotide score matrix, even when a biased mutational model is employed. Assuming that 30 bits are needed to distinguish a meaningful alignment from chance, this corresponds to a loss in sensitivity of over 11 bits, or a factor of over 2000. The loss is much greater if only match/mismatch scores are used.

While an alignment of two proteins diverged by fewer than 50 PAMs may be more significant when viewed using the nucleic acid alphabet, such an alignment in any case need be no longer than 15 residues to yield 30 bits of information. Thus, in the context of homology searches, it is generally only for noncoding regions that scores based simply on nucleic acid mutational models have real use.

TABLE 3

The Relative Information Available Using Protein- and Nucleic Acid-Based PAM Scores

Protein PAM distance	Information per residue (bits)	Nucleic acid PAM distance	Information per codon (bits)	Nucleic acid/ protein efficiency ratio
0	4.17	0	6.00	1.44
10	3.43	8	4.53	1.32
20	2.95	16	3.63	1.23
30	2.57	24	2.95	1.15
40	2.26	32	2.43	1.08
50	2.00	40	2.02	1.01
60	1.79	48	1.69	0.94
70	1.60	56	1.42	0.89
80	1.44	64	1.19	0.83
90	1.30	72	1.01	0.78
100	1.18	80	0.86	0.73
110	1.08	88	0.73	0.68
120	0.98	96	0.62	0.63
130	0.90	104	0.53	0.59
140	0.82	112	0.46	0.56
150	0.76	120	0.39	0.51

CONCLUSION

Nucleic acid sequence databases may be searched for a variety of reasons. To achieve optimal sensitivity it is necessary to use scores relevant to the specific question being asked. We have presented a variety of scoring matrices that can be used with BLASTP to implement such searches. For applications in which only alignments of nearly identical segments are of interest, a substantial improvement in sensitivity can be achieved by using PAM-5 scores.

The mutational model employed may significantly affect the scoring system, as illustrated by comparing the uniform substitution model to a mutational model in which transitions are more prevalent than transversions. In a search for conserved sequence elements in noncoding regions, the use of scores based on a biased as opposed to a uniform mutational model may substantially improve the search sensitivity. This is particularly important for alignments with <70% sequence identity, a range encompassing many mammalian and chordate noncoding homologous sequence elements. For coding sequence, scores based upon a model of amino acid substitutions (5, 7) will be superior for all but the closest relationships.

REFERENCES

1. Smith, T. F., and Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
3. Li, W. H., Wu, C. I., and Luo, C. C. (1985) *Mol. Biol. Evol.* **22,** 150–174.
4. Wilson, A. C., Carlson, S. S., and White, T. J. (1977) *Annu. Rev. Biochem.* **46,** 573–639.
5. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) *in* Atlas of Protein Sequence and Structure (Dayhoff, M. O., Ed.), Vol. 5, Suppl. 3, pp. 345–352, Natl. Biomed. Res. Found., Washington, DC.
6. Fitch, W. M., and Margoliash, E. (1967) *Science* **155,** 279–284.
7. Altschul, S. F. (1991) *J. Mol. Biol.* **219,** 555–565.
8. Karlin, S., and Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
9. Tuck, S. P., and Crawford, L. V. (1989) *Mol. Cell. Biol.* **9,** 2163–2172.
10. Bienz-Tadmor, B., Zakut-Houri, R., Libresco, S., Givol, D., and Oren, M. (1985) *EMBO J.* **4,** 3209–3213.