GENOMICS **46,** 24–36 (1997)
ARTICLE NO. GE974995

Comparison of DNA Sequences with Protein Sequences

William R. Pearson,*[,1] Todd Wood,* Zheng Zhang,† and Webb Miller†

*Department of Biochemistry, University of Virginia, Charlottesville, Virginia 22908; and †Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802

Received April 1, 1997; accepted August 25, 1997

The FASTA package of sequence comparison programs has been expanded to include FASTX and FASTY, which compare a DNA sequence to a protein sequence database, translating the DNA sequence in three frames and aligning the translated DNA sequence to each sequence in the protein database, allowing gaps and frameshifts. Also new are TFASTX and TFASTY, which compare a protein sequence to a DNA sequence database, translating each sequence in the DNA database in six frames and scoring alignments with gaps and frameshifts. FASTX and TFASTX allow only frameshifts between codons, while FASTY and TFASTY allow substitutions or frameshifts within a codon. We examined the performance of FASTX and FASTY using different gap-opening, gap-extension, frameshift, and nucleotide substitution penalties. In general, FASTX and FASTY perform equivalently when query sequences contain 0–10% errors. We also evaluated the statistical estimates reported by FASTX and FASTY. These estimates are quite accurate, except when an out-of-frame translation produces a low-complexity protein sequence. We used FASTX to scan the *Mycoplasma genitalium, Haemophilus influenzae,* and *Methanococcus jannaschii* genomes for unidentified or misidentified protein-coding genes. We found at least 9 new protein-coding genes in the three genomes and at least 35 genes with potentially incorrect boundaries.

INTRODUCTION

Advances in automated sequencing technologies have dramatically increased the rate of DNA sequence production and inclusion in the GenBank and EMBL DNA sequence databases. Indeed, although the traditional GenBank divisions have been growing at a relatively constant exponential rate, there have been dramatic increases in the amount of expressed sequence tag (EST) data over the past 2 years. For example, between Releases 81 (February 1994) and 99 (February 1996), the traditional divisions grew at a rate of 40% per year, from 163 million to 317 million bases, but the entire GenBank grew from 173 million to 463 million bases, because of the growth of the EST database. At the end of 1996, about 40% of the bases in GenBank were determined by high-throughput EST or genomic sequencing.

The DNA sequences produced by single-pass EST sequencing and high-throughput sequencing may be of lower quality than traditional "finished" GenBank sequences, which are typically based on multiple sequence reads from both strands of the DNA template. As a result, EST sequences are more likely to contain errors that produce frameshifts when translated into protein. Frameshift errors can be especially troublesome in searches with single-pass EST sequences, because these sequences are very likely to contain protein-coding regions, which are much more effectively identified by protein, rather than DNA, sequence comparison.

In an earlier paper (Zhang *et al.*, 1997), we reported the development of rapid algorithms for comparing a translated DNA sequence to a protein sequence within a band, for incorporation into the optimization stage of the FASTA program (Pearson, 1990), and of a full Smith–Waterman translated-DNA–protein alignment. In this paper, we consider two methods for aligning a translated-DNA sequence to a protein sequence and evaluate how well the two approaches identify distantly related sequences in the presence of DNA sequence errors. We also examine the accuracy of statistical estimates produced for translated-DNA–protein sequence alignment. The estimates can be quite accurate, but sometimes high scores are produced between unrelated sequences because of simple-sequence, low-complexity regions that are produced by translating the incorrect reading frames. This problem can be avoided by searching databases from which simple-sequence regions have been removed with **seg** (Wootton, 1994).

MATERIALS AND METHODS

Sequence databases. Sequence similarity searches were performed using a slightly modified version (PIR39b) of the annotated PIR39 database described earlier (Pearson, 1995). The database was modified

[1] To whom correspondence should be addressed at the Department of Biochemistry, Jordan Hall Box 440, University of Virginia, Charlottesville, VA 22908. Telephone: (804) 924-2818. Fax: (804) 924-5069. E-mail: wrp@virginia.EDU.

0888-7543/97 $25.00

to join some clearly related superfamilies (W.R.P, manuscript in preparation; the PIR39b database is available for downloading from ftp. virginia.edu:/pub/fasta). Two sequences from each of 46 families of proteins were used for these tests. The cDNA sequences, and their corresponding open reading frames, encoding these 92 sequences were identified in GenBank and used to evaluate DNA–protein sequence comparison with FASTX and FASTY. The annotations in the PIR39b database allow us to identify all of the sequences in the database that are related to members of the 46 query sequence families and, conversely, to identify the highest scoring unrelated sequences.

The cDNA sequences that encode the 92 query protein sequences, and their corresponding open reading frames (ORFs), were mutated to simulate the errors encountered in high-throughput EST sequencing. Based on the past experience of the Washington University St. Louis EST sequencing project (Hillier *et al.*, 1996), the frequency of errors (substitutions, insertions, and deletions) in high-throughput EST sequencing ranges from about 2.5 to 7.5%. Mutations were created at random locations at approximately 1, 2, 5, or 10% of the positions, with the total number of mutations distributed among deletions, insertions, and substitutions according to the ratio 1:1.54:1.86. This ratio is derived from observed discrepancies between 5′ ESTs and mRNA sequences (Fig. 3 and Table 6 in Hillier *et al.*, 1996; Table 6 has the labels for insertions and deletions reversed, L. Hillier, pers. comm.). Once these percentages are calculated, a normally distributed random number with mean and variance equal to the number of each type of change is drawn, and that number of substitutions, insertions, and deletions are introduced at random positions.

Comparison of search performance. Scoring parameters and comparison algorithms (FASTX, FASTY) were compared using methods described earlier (Pearson, 1995). Briefly, an "equivalence number" (the number of related sequences missed at a similarity score that balances the number of related sequences at or below the score with the number of unrelated sequences above) is calculated for searches with each of the 92 query sequences for each combination of gap penalties and frameshift penalties or for each algorithm. The performance of one search condition is compared to that of another by comparing the equivalence numbers for the 92 searches and recording a + or − depending on which search condition performed better (ties are ignored). The sign test is then used to determine if the distribution of +'s and −'s is significantly different from the null hypothesis that differences in performance are the result of random variation. Differences in performance are summarized by a "*z* value"; comparisons with *z* values >2 are statistically significant at the 0.05 level.

Characterization of bacterial genomes. To evaluate the boundaries of genes annotated for the *Mycoplasma genitalium, Haemophilus influenzae,* and *Methanococcus jannaschii* genomes, ORF sequences and their flanking nucleotides were extracted from the appropriate genome sequence. For the *M. jannaschii* genome, this was done for ORFs that differed in length by at least 20 amino acid residues when compared to homologues detected by a FASTA search of the translated ORFs (provided by The Institute for Genome Research; TIGR). ORFs and their flanking sequences were extracted from the complete *M. jannaschii* genome using the program **ext.** As input, **ext** requires a file containing a list of ORF names and positions. **ext** produces files containing original ORFs as well as files containing the ORFs with 250, 500, and 1000 flanking nucleotides at both 5′ and 3′ ends (for a total of 500, 1000, and 2000 additional nucleotides). These sequences were searched against Swiss-Prot using FASTX, and the results were examined for a combination of increase in alignment length with a decrease in expectation value when flanking nucleotides were included. Sequences that met these criteria were individually examined for the presence of low-complexity matches.

A second method of detecting incorrect ORF boundary assignments was also employed for these bacterial genomes. Sequences from regions greater than 60 nucleotides in length between ORFs were extracted and searched individually. This method also allows for the identification of ORFs that were not identified in the initial genomic analysis. These sequences were extracted using a modified version of **ext** called **exg,** which takes genome positions from command-line input. The FASTX search results were screened for matches that scored an expectation value of 0.01 or lower. Those matches were further screened individually both to omit low-complexity matches and to distinguish between previously unidentified ORFs and extensions of known ORFs.

```
A. FASTX
DNA sequence        AGTGTAGTGCCCTCTAGTTCA...
Frame 1              S  V  V  P  S  S  S
Frame 2               V  X  C  P  L  V
Frame 3                C  S  A  L  X  F
Protein sequence     S  V S  - P   G   S

                           *    *   *   *
B. FASTY
codonification     AGT GTA GT GCC CT CTAG TTCA
Protein sequence   S    V  S   -  P   G     S

                           *      *   *     *
```

FIG. 1. FASTX and FASTY alignments. An asterisk indicates a frameshift.

RESULTS

Two Notions of a DNA–Protein Alignment

We call the DNA sequence given as input the *determined DNA sequence,* to emphasize that it is determined by some experimental procedure and is subject to uncertainty. Both of our approaches to DNA–protein alignment determine, at least implicitly, a *hypothesized coding region,* or HCR, and align the conceptual translation of the HCR and the given amino acid sequence. The differences between the two approaches are confined to their methods of constructing an HCR.

The approach taken by FASTX can be described as follows. A *quasicodon* of the determined DNA sequence is any three consecutive nucleotides; quasicodons are numbered 2, 3, . . ., $N - 1$ according to their middle nucleotide, where N is the length of the determined sequence. An *allowable list of quasicodons* is any list that begins with quasicodon 2 and ends with quasicodon $N - 1$ and such that quasicodon $i < N - 1$ is always followed by quasicodon $i + 2$, $i + 3$, or $i + 4$. An allowable list of quasicodons determines the HCR obtained by concatenating the quasicodons one after the other. A FASTX alignment consists of an allowable list of quasicodons plus a protein alignment of the translated HCR and the given amino acid sequence.

Less formally, a FASTX alignment can be depicted as follows. Translate the determined DNA sequence in each of the three reading frames and place each amino acid below the central nucleotide of its quasicodon, as pictured in Fig. 1A. In essence, a FASTX alignment positions the entries of the original amino acid sequence underneath these translations, separating each entry by between one and three blanks, with the possible introduction of gaps.

A somewhat more general approach to DNA–protein alignments is taken by FASTY. A *codonification* of a nucleotide sequence $B = b_1b_2 . . . b_N$ is a sequence $c_1c_2 . . . c_L$, where (i) each c_k is either a nucleotide sequence

```
True coding region:          ..AGT G TA GCC..
Determined coding region:    ..AGT GCTA GCC..
HCR 1:                       ..AGT  CTA GCC..
HCR 2:                       ..AGT GCT  GCC..
```

FIG. 2. Two choices of hypothesized codon for FASTX. A "C" has been inserted between the first and the second positions in the second of three codons. FASTX is restricted to picking either "CTA" or "GCT" for its HCR. FASTY is able to pick the true codon.

with $2 \leq |c_k| \leq 4$ or a single dash character ($|c_k|$ denotes the length of c_k) and (ii) concatenating all the non-dash c_k sequences, in that order, gives B. A FASTY alignment between a determined DNA sequence B and an amino acid sequence A consists of a codonification, $c_1c_2 \ldots c_L$, of B with each c_k placed above an amino acid or a dash character, and where (1) if c_k appears over a dash, then $|c_k| = 3$ and (2) removing all dashes from the second row gives A. Figure 1B depicts a FASTY alignment.

In a FASTY alignment, each non-dash c_k corresponds to a codon of the HCR according to the following rules. If c_k is deleted (i.e., appears over a dash), then the codon is just c_k. Otherwise, letting a denote the amino acid aligned to c_k, the codon c is chosen to maximize the BLOSUM50 score of (the conceptual translation of) c and a minus the penalty for converting c_k to c. This conversion penalty deducts a fixed *frameshift penalty* for every nucleotide inserted or deleted and a fixed *miscall penalty* for every nucleotide substitution. In particular, a frameshift penalty is assessed whenever $|c_k|$ equals either 2 or 4.

In addition to directly capturing base miscalls (i.e., sequencing errors that cause an incorrect base to be reported), FASTY can correctly determine a wide range of frameshift errors caused by the sequencing process. In contrast, FASTX can determine only a more limited class of frameshift errors. With FASTX, insertion of an erroneous nucleotide must occur between two codons, and a nucleotide can be skipped (deleted) only if it occurs at both the last position of a codon and the first position of the subsequent codon. For instance, when a sequencing error inserts a nucleotide into a codon, FASTX is forced to pick one of the two codons obtained by dropping a base from one end or the other of the four-nucleotide sequence (assuming that FASTX correctly analyzes the codons on either side). See Fig. 2.

Computation of an optimal alignment of course presupposes that a score is assigned to every possible alignment. The score of a FASTX or FASTY alignment is defined as the score of the alignment between the translated HCR and the given amino acid sequence (assessed using BLOSUM-matrix substitution scores, plus gap-open and gap-extension penalties) minus penalties for the difference between the determined DNA sequence and the HCR. (Only frameshift penalties are assessed for FASTX.) FASTX and FASTY are guaranteed to compute an alignment that is optimal among all alignments of their respective types. Thus, the FASTY alignment will always score at least as high as the FASTX alignment. This fact alone does not guarantee that FASTY is superior to FASTX in all purposes; FASTY improves alignment scores for unrelated sequences as well as for homologous sequences, so under certain circumstances it can be less effective at distinguishing related sequences from unrelated sequences.

There are a variety of ways in which one might define an alignment of a DNA sequence and an amino acid sequence, some of which have been explored previously (Gish and States, 1993; Hein, 1994; Hein and Stovbaek, 1994; Knecht, 1995; Guan and Uberbacher, 1996; Huang and Zhang, 1996; Peltola *et al.*, 1986). In general, there is a tradeoff between (1) the completeness of the set of sequencing errors that are directly modeled by the underlying set of alignments and (2) the execution time required to compute an optimal alignment. The FASTX approach, which models only frameshift errors at codon boundaries, is similar to techniques developed by Guan and Uberbacher (1996), although judging by their published description, our algorithm (Zhang *et al.*, 1997) is more efficient than theirs. The basic idea behind FASTY was described by Peltola *et al.* (1986), though we have added a number of improvements, including use of modern protein-alignment scores, modeling of base miscalls, and implementation techniques (Zhang *et al.*, 1997) needed to make it competitive in execution time with FASTX. It is even possible to develop a rigorous alignment algorithm that, in some precise sense, models all possible sequencing errors (Zhang *et al.*, 1997), but its execution time is prohibitive.

Searching with FASTX

Programs for comparing translated DNA sequences to protein sequences, allowing frameshifts, have been added to the FASTA package (Pearson, 1996). FASTX and FASTY compare a DNA sequence to a protein sequence database, translating the DNA sequence in either the three forward or the three reverse frames. FASTX allows for frameshifts between codons, while FASTY allows for frameshifts within codons as well. The TFASTX and TFASTY programs compare a protein sequence to a DNA sequence database, translating each sequence in the database in three forward and three reverse frames.

FASTX, FASTY, TFASTX, and TFASTY use the same four steps that FASTA uses for calculating a similarity score: (1) using a lookup table to rapidly find regions with shared identical pairs of residues (*ktup = 2*) or single shared identities (*ktup = 1*), (2) rescanning the regions using a BLOSUM50 scoring matrix, (3) joining high-scoring regions that do not overlap, and (4) calculating an optimal Smith–Waterman score in a 16 (*ktup = 2*)- or 32 (*ktup = 1*)-residue-wide band (Chao *et al.*, 1992) centered around the best initial region. FASTX and FASTY differ from FASTA by using three protein sequences, from each of the three frames, for the initial lookup process (step 1). Step 3 is modified

(a) to cause high-scoring regions in adjacent reading frames to appear to be adjacent in the sequence, so they can be more easily joined, and (b) to allow small overlaps (10 residues) between joined regions. Step 4 is changed to produce a band-limited DNA–protein local alignment score (Zhang *et al.*, 1997) as outlined above. FASTX and FASTY use a full Smith–Waterman local DNA–protein alignment in linear space for the final alignment and alignment score (Zhang *et al.*, 1997).

TFASTX/Y use a similar strategy, but instead of augmenting the query-sequence lookup table, the library sequence is encoded as two separate three-frame translations, one forward and one reverse. Again, steps 3 and 4 are modified for DNA–protein comparison and TFASTX/Y provide a full Smith–Waterman alignment, without limits on gap size, for the final display.

Figures 3–5 show the output produced by a typical FASTX search. As with other programs in the FASTA package, the distribution of observed and expected similarity scores (Fig. 3) and the expectation value [*E*() value] of the highest scoring unrelated sequence (Fig. 4) can be used to evaluate the accuracy of the expectation values calculated by FASTX, FASTY, TFASTX, and TFASTY. In this search of the SwissProt database with a housefly glutathione transferase cDNA sequence, there is excellent agreement between the numbers of observed and expected similarity scores (Fig. 3), particularly for scores from 80 to 120 (all the scores >120 come from glutathione transferase sequences). Likewise, the expectation value for the highest scoring unrelated sequence should be ~1; in this example the highest scoring unrelated sequence has an *E*() of 1.6.

Figure 4 also shows the expectation values for a FASTA search of the same SwissProt database using the protein sequence as the query. In general, the related-sequence expectation values for the protein–protein comparison are about one-half the values calculated from the translated DNA–protein sequence comparison. The reduced statistical significance results from the increased probability of producing a high score from an unrelated sequence when three longer protein sequences (three translation frames that include a 3′ untranslated region) are compared. FASTX and FASTY compare either the forward or the reverse three-frame translation to the protein database; if both forward and reverse frames were compared, the statistical significance would be decreased another twofold because of the effectively larger number of comparisons performed.

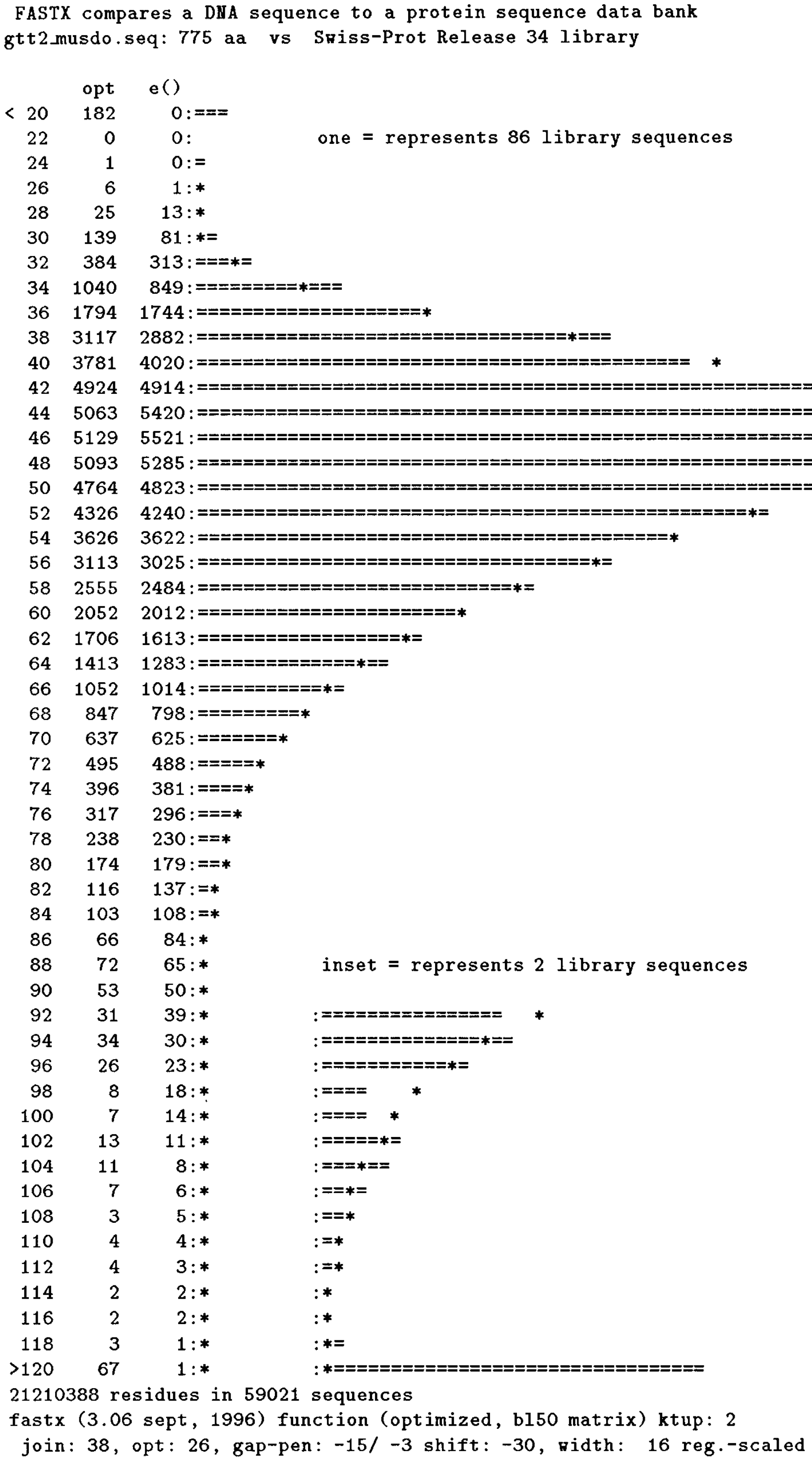
cDNA to protein sequence comparison can be surprisingly sensitive. The translated housefly glutathione transferase cDNA sequence shares statistically significant similarity with human and rodent class-theta enzymes (GTT1_HUMAN, GTT1_RAT), which must have diverged from a common ancestor more than 500 Myr ago, and with yeast URE2_YEAST glutamine repressor protein and various plant glutathione transferases (GTH3_ARATH, GTH3_MAIZE), which diverged more than 1 Byr ago, as well as several bacterial homologs (DCMA_METS1, DCMA_METSP, SSPA_HAEIN), which must have diverged more than 2 Byr ago. In contrast, DNA sequence searches with the housefly glutathione transferase cDNA against the appropriate GenBank divisions (primate/rodent, plant, or bacterial) do not find any of the mammalian, yeast, or bacterial homologues (*E*() > 10). Thus, FASTX/Y searches are only about two-fold less sensitive than FASTA searches with the encoded protein sequence, but dramatically more sensitive than the corresponding DNA sequence similarity search.

Figure 5 shows alignments between a mouse class-mu glutathione transferase, to which bases were added and deleted, and the correct GTM1_MOUSE protein sequence using FASTX (Fig. 5A) or FASTY (Fig. 5B). Both algorithms produce alignments that extend from the beginning of the protein sequence to the end, but the FASTY alignment does a better job of maximizing the number of identities (91.7% identity for FASTY, 59.6% for FASTX, but note that the *z* score, which indicates statistical significance, is slightly lower for the FASTY alignment).

Effective Search Parameters for FASTX and FASTY

Before evaluating the relative performances of FASTX and FASTY, we first searched an annotated PIR39 database to identify good combinations of gap-open, gap-extension, and frameshift penalties in the presence of sequence errors. Open reading frames from 92 query sequences from 46 protein families used in previous studies (Pearson, 1995) were "mutated" and used to evaluate search performance with different gap and frameshift penalties (Fig. 6). In general, high frameshift penalties are most effective when the error rate is 0, as expected, since no frameshifts should occur. However, even with modest amounts of errors (substitutions, insertions, and deletions) (1 and 2%), searches with frameshift penalties of −15 are significantly more effective than searches with higher penalties. Very similar results are seen when the entire cDNA sequence is used (data not shown). The error rates encountered in high-quality EST sequences are about 1.4% substitutions, 1.1% insertions, and 0.7% deletions (Hillier *et al.*, 1996, Table 6, corrected), so the 2–5% error rates shown in Figs. 6 and 7 are the most realistic.

FASTX and FASTY perform very similarly when the best scoring parameters are used (Fig. 7A). However, in general one will not know the error rate in advance and would prefer to use gap and frameshift penalties that perform well overall, for example, −15/−2/−20 for FASTX and −15/−2/−25/−30 for FASTY. With these penalties, FASTX performs a bit better on ORF sequences (Fig. 7A), otherwise the two programs perform about the same. Although some of the differences in performance are statistically significant, they are not very dramatic. Figure 7B reports the total number of related sequences "missed" in all 92 searches as a function of error rate. From this perspective, differences in performance between FASTX and FASTY, or even ORFs and cDNA sequences, are far less significant than changes in the error rate.



```
FASTX compares a DNA sequence to a protein sequence data bank
gtt2_musdo.seq: 775 aa  vs  Swiss-Prot Release 34 library

       opt   e()
< 20   182     0:===
  22     0     0:           one = represents 86 library sequences
  24     1     0:=
  26     6     1:*
  28    25    13:*
  30   139    81:*=
  32   384   313:===*=
  34  1040   849:=========*===
  36  1794  1744:====================*
  38  3117  2882:=================================*===
  40  3781  4020:============================================  *
  42  4924  4914:=========================================================*
  44  5063  5420:===========================================================*
  46  5129  5521:===========================================================*
  48  5093  5285:===========================================================*
  50  4764  4823:========================================================*
  52  4326  4240:=================================================*=
  54  3626  3622:==========================================*
  56  3113  3025:===================================*=
  58  2555  2484:============================*=
  60  2052  2012:=======================*
  62  1706  1613:==================*=
  64  1413  1283:==============*==
  66  1052  1014:===========*=
  68   847   798:=========*
  70   637   625:=======*
  72   495   488:=====*
  74   396   381:====*
  76   317   296:===*
  78   238   230:==*
  80   174   179:==*
  82   116   137:=*
  84   103   108:=*
  86    66    84:*
  88    72    65:*        inset = represents 2 library sequences
  90    53    50:*
  92    31    39:*        :================   *
  94    34    30:*        :==============*==
  96    26    23:*        :===========*=
  98     8    18:*        :====    *
 100     7    14:*        :====  *
 102    13    11:*        :=====*=
 104    11     8:*        :===*==
 106     7     6:*        :==*=
 108     3     5:*        :==*
 110     4     4:*        :=*
 112     4     3:*        :=*
 114     2     2:*        :*
 116     2     2:*        :*
 118     3     1:*        :*=
>120    67     1:*        :*=================================
21210388 residues in 59021 sequences
fastx (3.06 sept, 1996) function (optimized, bl50 matrix) ktup: 2
 join: 38, opt: 26, gap-pen: -15/ -3 shift: -30, width:  16 reg.-scaled
```

FIG. 3. Distribution of FASTX similarity scores. A housefly glutathione transferase cDNA sequence (GenBank Accession No. X73574) was used to search the SwissProt protein database (Release 34) using the FASTX program. The BLOSUM50 matrix (Henikoff and Henikoff, 1992) was used with a penalty of −15 for the first residue in a gap, −3 for each additional residue, and −30 for a frameshift. "= = =" denotes the number of sequences in the database obtaining the similarity score shown; "*" indicates the number of sequences expected to obtain a similarity score.

Statistical Estimates from DNA–Protein Comparisons

FASTX/Y and TFASTX/Y also report estimates of the statistical significance of the similarity scores; these estimates are based on the relationship between the mean unrelated sequence similarity score and library sequence and the average variance of the unrelated scores (W.R.P, in preparation). For protein sequences, the statistical estimates are quite accurate; on average about 1/10th of the sequences have expectation values with probabilities ≤0.1, 1/2 have expectation values

			FASTX		FASTA
The best scores are:		initn	opt	E(58,772)	(58,763)
GTT2_MUSDO	GLUTATHIONE S-TRANSFERASE 2	1414	1414	0	0
GTT1_LUCCU	GLUTATHIONE S-TRANSFERASE 1-1	1020	1035	0	0
GTA_PLEPL	GLUTATHIONE S-TRANSFERASE A	222	256	4e-13	2.1e-13
GTH3_ARATH	GLUTATHIONE S-TRANSFERASE ERD13	243	243	3.8e-12	2.0e-12
GTT1_HUMAN	GLUTHATHIONE S-TRANSFERASE THETA	214	243	4.2e-12	2.2e-12
GTT1_RAT	GLUTHATHIONE S-TRANSFERASE 5	215	233	2.5e-11	1.3e-11
DCMA_METS1	DICHLOROMETHANE DEHALOGENASE	118	198	1.3e-08	1.9e-08
GTT2_DIACA	GLUTATHIONE S-TRANSFERASE 2	158	195	1.4e-08	1.2e-08
GTH3_MAIZE	GLUTATHIONE S-TRANSFERASE III	194	178	3.9e-07	2.0e-07
URE2_YEAST	URE2 PROTEIN.	133	179	4.9e-07	2.5e-07
GTH_HYOMU	GLUTATHIONE S-TRANSFERASE	137	176	5.4e-07	2.7e-07
DCMA_METSP	DICHLOROMETHANE DEHALOGENASE	123	166	4.1e-06	5.0e-06
GT_PROMI	GLUTATHIONE S-TRANSFERASE GST-6.0	41	152	3.7e-05	1.8e-05
GTH1_ARATH	GLUTATHIONE S-TRANSFERASE ERD11	122	144	0.00015	7.5e-05
SSPA_ECOLI	STRINGENT STARVATION PROTEIN A.	111	143	0.00019	9.1e-05
GT_ECOLI	GLUTATHIONE S-TRANSFERASE	35	141	0.00026	0.00012
EF1G_HUMAN	ELONGATION FACTOR 1-GAMMA	37	143	0.00034	0.00017
SSPA_HAESO	STRINGENT STARVATION PROTEIN A	93	137	0.00054	0.00031
EF1G_RABIT	ELONGATION FACTOR 1-GAMMA	37	139	0.00069	0.00035
GT_HAEIN	GLUTATHIONE S-TRANSFERASE	104	127	0.0032	0.0015
GTT1_CHICK	GLUTATHIONE S-TRANSFERASE 1 (EC	152	128	0.0032	0.0015
LGUL_SOYBN	LACTOYLGLUTATHIONE LYASE	89	126	0.0039	0.0019
GTY2_ISSOR	GLUTATHIONE S-TRANSFERASE Y-2	66	125	0.0042	0.002
SSPA_HAEIN	STRINGENT STARVATION PROTEIN A	92	125	0.0046	0.0031
LIGF_PSEPA	LIGF PROTEIN.	103	121	0.011	0.0089
GTX1_NICPL	PROB. GLUTATHIONE S-TRANSFERASE	40	113	0.039	0.019
EF1G_SCHPO	ELONGATION FACTOR 1-GAMMA	40	111	0.094	0.046
GTXA_ARATH	GLUTATHIONE S-TRANSFERASE	31	107	0.12	0.055
GTH_BRAOL	GLUTHATHIONE S-TRANSFERASE	105	94	0.47	0.21
EF1B_HUMAN	ELONGATION FACTOR 1-B	92	96	0.81	0.38
*RRNA_YEAST	RNA POLYMERASE I SPEC TRANS	68	90	1.6	> 10
EF1B_RABIT	ELONGATION FACTOR 1-B	85	92	1.6	0.78
*Y279_HAEIN	HYPOTHETICAL PROTEIN HI0279	47	87	2	> 10
GT27_SCHMA	GLUTATHIONE S-TRANSFERASE 26 KD	42	90	2.3	1.1
*RAFD_ECOLI	RAFFINOSE INVERTASE	68	93	2.6	1.3
*SYEP_HUMAN	MULTIFUNCTION	77	98	2.7	1.3
GTMU_RABIT	GLUTATHIONE S-TRANSFERASE MU 1	90	89	2.7	1.3
*TYPH_ECOLI	THYMIDINE PHOSPHORYLASE	40	91	3.4	> 10
*VE2_HPV24	REGULATORY PROTEIN E2.	66	90	4.3	> 10
*YQJG_ECOLI	HYPOTHETICAL 37.4 KD PROTEIN	28	88	4.6	2.2

FIG. 4. FASTX search—high scoring sequences. FASTX similarity scores from the search in Fig. 3 for high-scoring related and *unrelated* protein sequences are shown. Unrelated sequences are highlighted in italics. The statistical significance, in the form of an *E*() for each similarity score is shown (FASTX). In addition, expectation values for the same search with the translated GTT2_MUSDO protein sequence are shown (FASTA).

with probabilities ≤0.5, and so on. Statistical estimates for FASTX searches are sometimes less accurate, because out-of-frame translations can sometimes have low-complexity amino acid sequence runs that produce statistically significant similarity scores with similar regions in the protein database. Thus, when 90 ORF sequences were shuffled to produce random sequences, and these sequences were used to search PIR39b, 6 random query sequences found a highest scoring alignment with expectation values from $<10^{-8}$ to 0.0005. These values are much lower than expected by chance, and six more had $E() < 0.05$. Likewise, when unshuffled ORF sequences were used, 6 queries had highest scoring unrelated sequences with $E() < 10^{-20}$, and 20 of 90 sequences had $E() < 0.05$ for the highest scoring unrelated sequence. However, about half of the query

A. FASTX alignment

```
>>GTM1_MOUSE GLUTATHIONE S-TRANSFERASE GT8.7 (EC 2.5.1.18) (GST 1-1) (CLASS-M (217 aa)
 initn: 488 init1: 395 opt: 1120 Z-score: 1290.7 expect()    0
Smith-Waterman score: 1120;  59.565% identity in 230 aa overlap

              50        60        70        80        90       100       110       120
mGSTM1 PMI/MGYWKVRGLTHPIRMLLEYTDPSYDEKRYTMGDG\PDFDRSQWLNEKFKLGLEF\PNLPYLIDGSHKITQRMPS/L
       ::: .:::.:::::::::::::::: :::::::::::. :::::::::::::::::.: :::::::::::::::     :
GTM1_M PMI-LGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDA-PDFDRSQWLNEKFKLGLDF-PNLPYLIDGSHKITQSNAI-L
                10        20        30         40        50         60        70

             130       140       150       160       170       180       190       200
mGSTM1 RYLATKPT/LEEMTEEERIRADIVENQIAWKPA\QLIMLSLQPXLX\KQKPEFLKTIPEKMSSTLSSW/GKRPWFAWDKC
       :::: :   :.  ::::::::::::::.      :::::  .: .  ::::::::::::::.       ::::::: ::
GTM1_M RYLARKHH-LDGETEEERIRADIVENQVMDTRM-QLIMLCYNPDFE-KQKPEFLKTIPEKMKLYSEFL-GKRPWFAGDKV
         80         90       100        110       120        130       140        150

             210       220       230       240       250       260       270
mGSTM1 HLCGFLC\YDILDQYRMFEPSAWTPSQTX/RXLPGPLRGPQ\KISALHEEXPVHRHS\HIYKDGPLEXQA
           ::  ::::::::::::.      .  : . . ..: . ::::  .          . : .    .
GTM1_M TYVDFLA-YDILDQYRMFEPKCLDAFPNL-RDFLARFEGLK-KISAYMKSSRYIATP-IFSKMAHWSNK
             160       170       180        190        200        210
```

B. FASTY alignment

```
>>GTM1_MOUSE GLUTATHIONE S-TRANSFERASE GT8.7 (EC 2.5.1.18) (GST 1-1) (CLASS-M (217 aa)
 initn: 488 init1: 395 opt: 1099 Z-score: 1286.6 expect()    0
Smith-Waterman score: 1099;  91.705% identity in 217 aa overlap

              50        60        70        80        90       100       110       120
mGSTM1 PMI\LGYWKVRGLTHPIRMLLEYTDPSYDEKRYTMGD/APDFDRSQWLNEKFKLGL/DFPNLPYLIDGSHKITQ\SNAIL
       ::: ::::.:::::::::::::::: ::::::::::: :::::::::::::::::: ::::::::::::::::: :::::
GTM1_M PMI LGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGD APDFDRSQWLNEKFKLGL DFPNLPYLIDGSHKITQ SNAIL
                10        20        30         40        50         60        70

             130       140       150       160       170       180       190       200
mGSTM1 RYLATK\HHLEEMTEEERIRADIVENQIA/XNPLQLIML/CYNLDFEKQKPEFLKTIPEKM\KLYSEFLGKRPWFAWDK/
       :::: : :::.  ::::::::::::::.   . .::::: ::: ::::::::::::::::: :::::::::::::: ::
GTM1_M RYLARK HHLDGETEEERIRADIVENQVM DTRMQLIML CYNPDFEKQKPEFLKTIPEKM KLYSEFLGKRPWFAGDK
         80         90       100        110        120       130        140       150

             210       220       230       240       250       260       270
mGSTM1 VTYVDFFAYDILDQYRMFE\PKCLDAFPNLR/DFLARFEGLKKISA/YMKSSRYIGTA/IFTKMAHWSNK
       ::::::.:::::::::::: ::::::::::: :::::::::::::: ::::::::.:  ::.::::::::
GTM1_M VTYVDFLAYDILDQYRMFE PKCLDAFPNLR DFLARFEGLKKISA YMKSSRYIATP IFSKMAHWSNK
             160       170        180        190        200        210
```

FIG. 5. Translated DNA–protein alignments—FASTX and FASTY. Alignments between mGSTM1.e05, a modified copy of the GTM1_MOUSE cDNA sequence with ~5% mutations (mGSTM1.e05 is 94.2% identical to the correct cDNA), and its encoded protein sequence, using FASTX and FASTY. The BLOSUM50 matrix was used with a penalty of −15 for the first residue in a gap, −2 for each additional residue, −15 for a frameshift, and −15 for a nucleotide substitution (FASTY only).

sequences had $E() > 0.5$ for the highest scoring unrelated sequences, suggesting that for many of these sequences, the statistical estimates were accurate. The accuracy of the statistical estimates can be judged by a quantile–quantile plot (Fig. 8). The expectation value of the highest unrelated sequence score from each of 90 ORF query sequences (or 90 random sequences derived from the 90 ORF sequences) was sorted from lowest value to highest, converted to a Poisson probability value using the equation $P(E) = 1 - e^{-E}$, and plotted against the cumulative fraction of sequences examined. If there is perfect agreement between the probability of a high score and the number of times it occurs, then the points will fall on a diagonal line with a slope of 1. If many of the points fall above the diagonal, then there are fewer sequences with low probabilities than expected and thus the estimates are conservative. If many points fall below the diagonal, as occurs in Fig. 8A, then low probabilities have been assigned too frequently, and a calculated "probability" of 0.001 (or even lower) happens far more often than the expected 0.1% of the time. Thus, a very low expectation value is not necessarily statistically significant.

Examination of the high-scoring, statistically significant FASTX alignments showed that in every case, FASTX had produced an out-of-frame translation that

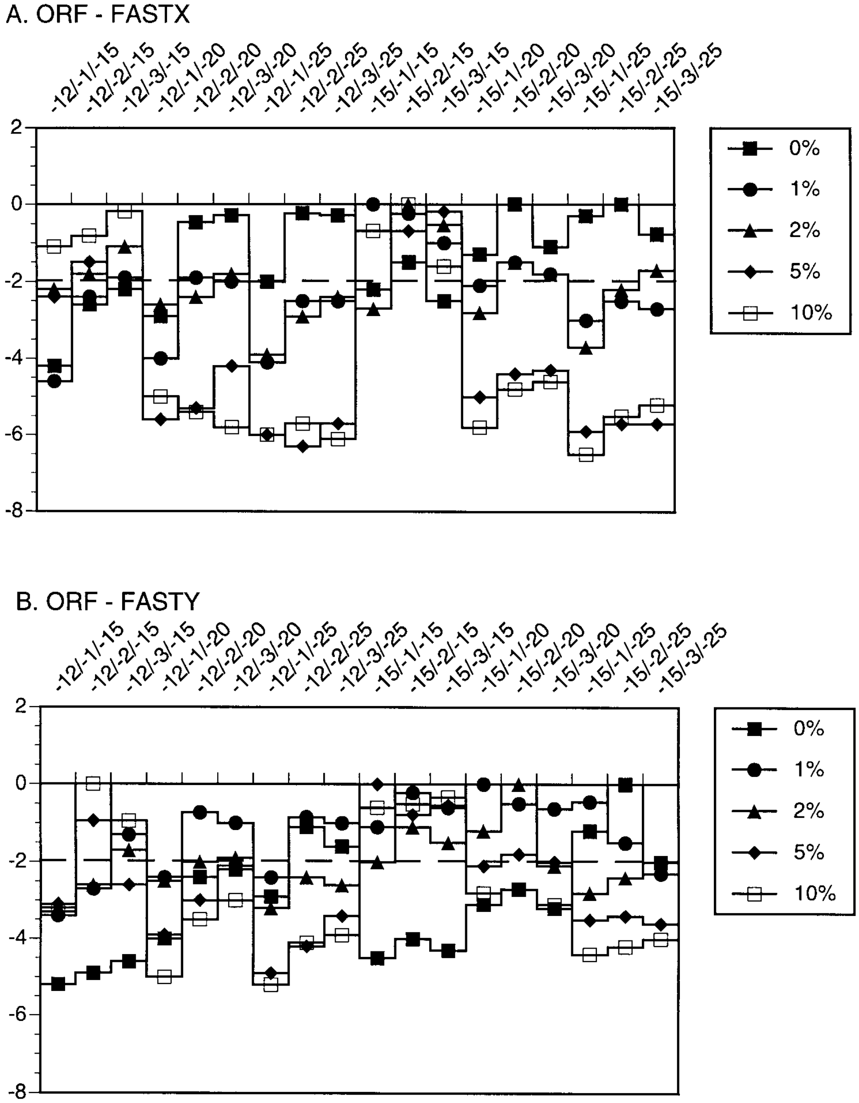


FIG. 6. Effective scoring penalties—open reading frames. The *z* values for the difference in performance with respect to the best performing gap and frameshift penalties. The values across the top (e.g., −12/−1/−15) indicate the cost of the first residue in a gap, each additional gapped residue, and a frameshift. The first-gap value is slightly different from the penalties defined in the description of the alignment algorithms. Thus, a first-gapped residue cost of −12 and a gap extension cost of −1 are equivalent to a gap open penalty of 11 and a gap extension penalty of 1. Differences in performance are presented as *z* values for a sign test of the equivalence numbers; *z* values greater than 2 are significant at the 0.05 level. Positive *z* values indicate that the parameters at the top of the column performed better; negative values indicate that the best parameter combination for that error rate performed better. (**A**) Searches with FASTX. The best parameters for FASTX were −15/−2/−20, −15/−1/−15, −15/−2/−15, −15/−1/−15, and −15/−2/−15, for 0, 1, 2, 5, and 10% errors in ORF sequences. (**B**) Searches with FASTY. All the searches shown used a substitution cost of −30. Searches done with a substitution cost of −15 were no better, but not significantly different. The best FASTY parameters on ORFs were −15/−2/−25/−30 (the last value is the substitution cost), −15/−1 /−20/−30, −15/−2/−20/−30, −15/−2/−15/−30, and −12/−2/−15/−30.

yielded a protein "domain" with a repetitive, low-complexity sequence. To evaluate the statistical estimates in the absence of these low-complexity matches, the **seg** program (Wootton, 1994) was used to strip the PIR39b database of these regions (they were replaced by "x's"). When low-complexity regions are removed from the protein database, the statistical estimates calculated both for random sequences and for the highest scoring unrelated sequence are reliable and somewhat conservative (Fig. 8B). Thus, FASTX can calculate accurate statistical significance estimates for the local similarity scores if low-complexity regions are removed from the protein sequence database. However, when one scans an unmodified database, scores that appear significant should be examined carefully for alignment of low-complexity regions.

TFASTX/Y, an Alternative to TFASTA

The original FASTA package (Pearson and Lipman, 1988) provided TFASTA, a program to compare a protein sequence to a DNA sequence library, calculating six scores, one for each of the three forward and the three reverse reading frames. TFASTX and TFASTY

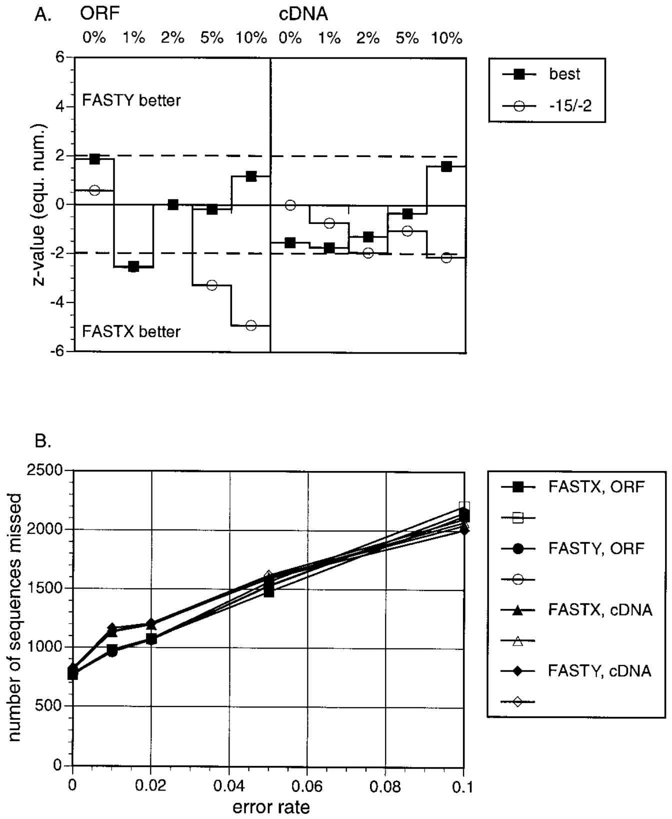


FIG. 7. FASTX versus FASTY. (**A**) The best scoring parameters for FASTX and FASTY searches using either ORF or cDNA query sequences were compared for each error rate "best." Differences in performance are presented as z values; positive z values indicate that FASTY performed better; negative values support FASTX. The best ORF parameters are shown in Fig. 6. For cDNAs, the best penalties were −15/−1/−25, −15/−2/−15, −15/−2/−15, −15/−2/−15, and −12/−3/−15 (FASTX) and −15/−1/−25/−30, −15/−2/−25/−30, −15/−1/−20/−30, −15/−3/−15/−30, and −12/−3/−15/−30 (FASTY). Searches with a good general purpose combination of parameters, −15/−2/−20 for FASTX and −15/−2/−25/−30 for FASTY, are also shown. (**B**) The total number of related sequences missed using FASTX or FASTY with error rate. Filled symbols show the results of searches using the best search parameters for the error rate; open symbols show the results with the general purpose parameters used in (A).

use the same DNA–protein alignment algorithms as FASTX and FASTY to provide two alignment scores for each DNA sequence; one from the forward and one from the reverse-complement sequence. TFASTA statistics are based on the best of the six scores produced from each library sequence; TFASTX statistics are based on the best of the forward and reverse similarity scores. In the example shown in Fig. 9, TFASTX does not perform substantially better than TFASTA in identifying distantly related DNA sequences, although both TFASTA and TFASTX perform substantially better than a DNA–DNA comparison. However, TFASTX provides much more informative alignments when sequencing errors or short introns are present.

A quick glance at the names of the protein sequences in Fig. 9 suggests that many "significant" scores were calculated for unrelated sequences (e.g., S75161 ribosomal protein L13, $E() < 0.0012$; GMGLYO glyoxylase, $E() < 0.0051$; SMAPHOAA alkaline phosphatase, $E() < 0.03$). However, additional searches with these DNA sequences using FASTX revealed that they share strongly significant similarity with many members of the glutathione transferase family, and thus are likely to be homologous to the housefly query sequence. (SMAPHOAA clearly encodes an alkaline phosphatase starting at nucleotide 667, but also encodes a glutathione transferase family member from residue 2 to 511.) The highest scoring sequences in the GenBank divisions used for this example that are clearly unrelated to GTT2_MUSDO are PSEPUOH109 [$E() < 0.39$], which encodes a haloacetate dehalogenase, and BMOEF1BP [$E() < 0.57$], which encodes a EF1β elongation factor that does not appear to be related to the EF1γ elongation factors that are homologous to glutathione transferases (Koonin *et al.*, 1994). Thus, for this query sequence, which does not contain low-complexity regions, the statistical estimates appear quite accurate. A comprehensive evaluation of the statistics of comparisons to DNA databases will require a carefully annotated DNA sequence database.

Identification of Genes in Bacterial Genome Sequences

FASTX and FASTY can also be used to analyze genomes that lack long exons in protein-coding genes. As of this writing, full genomic sequences of six prokaryotes and one eukaryote are available. Here, we present results of preliminary analyses of three prokaryotic genomes sequenced at TIGR: *M. jannaschii* (Bult *et al.*, 1996), *H. influenzae* (Fleischmann *et al.*, 1995), and *M.*

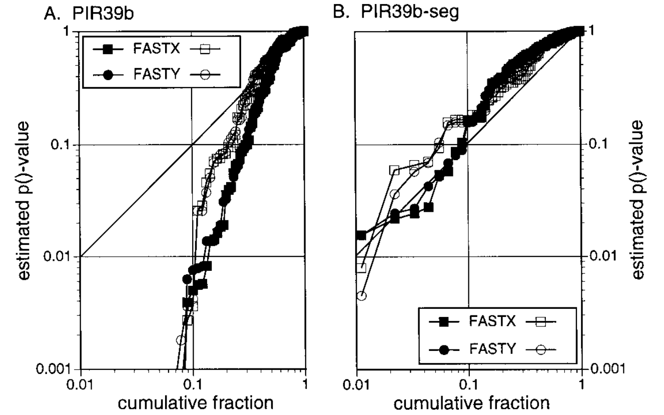


FIG. 8. Statistics of FASTX and FASTY scores. Probability of the score of the highest scoring unrelated sequence versus the cumulative fraction of sequences examined. Searches were performed with 90 ORF query sequences (open symbols) or 90 random sequences (filled symbols) derived from ORF sequences against either the unmodified PIR39b database (**A**) or the PIR39b database with low-complexity sequences removed (**B**). Searches were performed with the three parameter sets shown. As many as 7 sequences with $P()$ values <0.001 (some $<10^{-20}$) of the 90 query sequences used for each test set are below the bottom of the plot in A. FASTX searches used parameters of −15/−2/−20; FASTY searches used −15/−2/−25/−30.

				TFASTX			TFASTA
The best scores are:				initn	opt	E(161,671)	(161,442)
MDGST2A	M. domesticus GST	775	[f]	1414	1414	0	0
DMGST	D. melanogaster GST1-1	764	[f]	1009	1021	0	0
LUCGLTR	L. cuprina GST	790	[f]	992	1007	0	0
AGGST15	A. gambiae GST1-5	748	[f]	899	918	0	0
MOTGLUSTRA	M. sexta GST	951	[f]	307	433	1e-28	5.2e-30
ATHERD13	A. thaliana GST	836	[f]	244	243	7.5e-13	7.8e-14
CSU42463	Coccomyxa GST	1252	[f]	173	242	1.2e-12	1.2e-13
HSGSTT1	H. sapiens GSTT1	1004	[f]	261	231	8.7e-12	7.6e-14
DCCARSR8	D. caryophyllus CARSR8	957	[f]	209	229	1.2e-11	5.5e-13
HUMGSTT2A	H. sapiens GSTT2	1036	[f]	196	219	9e-11	1.2e-11
MYMDCMA	dichloromethane dehalogenase	2283	[f]	140	192	3e-08	4.2e-09
ZMGST3	Maize GST	913	[f]	194	178	2.2e-07	4.3e-08
YSCURE2	S. cerevisiae URE2	1427	[f]	154	179	2.6e-07	5.5e-08
ECAE000186	E. coli genome	9963	[r]	82	176	2e-06	4.7e-07
D90862	E. coli genome	17036	[f]	148	175	3.6e-06	2.9e-07
SYCCPNC	Synechocystis genome, 20/27,	79780	[r]	175	177	7.8e-06	2.3e-06
PSEAA	P. aeruginosa gene	392	[r]	100	150	2.6e-05	3.5e-06
MTBDCMAA	dichloromethane DH	5720	[f]	101	161	2.3e-05	1.5e-05
ATU70672	A. thalia GST	760	[f]	156	152	2.9e-05	4.5e-06
EGU80615	E. globulus auxin-induced prot.	841	[f]	80	148	6.8e-05	2.1e-05
PMU38482	P. mirabilis gstB	2123	[f]	59	152	6.3e-05	3.5e-05
S43311	A. aegypti GST	134	[f]	99	137	0.00014	3.0e-05
TOBAPI2B	Tobacco api2 mRNA	892	[f]	154	142	0.00023	4.9e-05
ECSSPG	E. coli stringent starvation prot.	1616	[f]	111	143	0.00029	4.0e-05
NTC7	N.tabacum mRNA C-7. 796	951	[f]	60	140	0.00035	0.00012
S75161	rplM=ribosomal protein L13	1842	[f]	93	136	0.0012	0.00042
ECAE000402	E. coli genome	10713	[r]	155	143	0.0012	0.00027
TOBPARB	N. tabacum GST	913	[f]	154	132	0.0016	0.00021
NTAUX107	N. tabacum auxin-induced	941	[f]	74	130	0.0024	0.0011
NTPARC	N. tabacum parC mRNA. 796	989	[f]	93	128	0.0036	0.0017
ZMGST27	Z. mays GST-27	954	[f]	119	127	0.0043	0.0011
ZMU12679	Z. mays GST	985	[f]	119	127	0.0044	0.0011
GMGLYO	G.max mRNA for Glyoxalase	932	[f]	89	126	0.0051	0.0011
TRBTCAC2X	T. cruzi stress/GST	1618	[f]	84	122	0.017	0.014
SMAPHOAA	S. marcescens alk. phosph.	2672	[f]	82	121	0.03	0.0068
HIU32696	H. influenzae genome	10574	[f]	140	127	0.026	0.027
S44036	msr-1=multiple stimulus re	906	[f]	40	116	0.034	0.014
PSELIG	paucimobilis b-etherase	3044	[f]	103	118	0.058	0.032
HIU32822	H. influenzae genome	14027	[r]	92	123	0.071	0.044
*PSEPUOH109	Plasmid pUOH109	2231	[f]	113	107	0.39	0.064
*BMOEF1BP	Bombyx mori mRNA	794	[f]	42	101	0.57	0.14
SCCH13LST	S.cerevisiae chromosom	37396	[r]	134	113	1	0.0001

FIG. 9. TFASTX—high-scoring sequences. High-scoring sequences is a search of the Primate, Invertebrate, Plant, and Bacterial sections of GenBank (Release 99) using a housefly glutathione transferase (GTT2_MUSDO). Unrelated sequences are highlighted in italics. (In cases in which the sequence was not clearly labeled as a glutathione transferase, additional searches were done with FASTX or FASTA against the SwissProt protein sequence database, and homology was established if the GenBank sequence obtained *E*() values $<10^{-6}$ with many glutathione transferase family members. Also shown are the expectation values calculated from a TFASTA search.

genitalium (Fraser *et al.*, 1995). The sequences were downloaded via ftp from TIGR's World Wide Web site (http://www.tigr.org/). These genomes were chosen for this preliminary study strictly because of the similarity of their on-line documentation.

Identification of protein-coding regions of genomic

TABLE 1

Modified Gene Boundaries Based on Extended Alignments

		ORF		Extended ORF			
Name	Match	*E*()	Length	*E*()	Length	Start	Stop
Haemophilus influenzae							
HI0097	FBP_HAEIN	0	294	0	333	103688	104682
HI0117	MLTA_ECOLI	2.2 e −17	81	0	269	131566	132369
HI0153	DCUB_ECOLI	1.1 e −07	58	0	436	170137	168837
H10187	YIGT_ECOLI	6.2 e −08	48	7.9 e −20	254	200711	201501
HI0200	SELD_HAEIN	0	322	0	348	215291	216328
HI0218	T1R1_ECOLI	0	160	0	1032	234851	237908
HI0220	ARCB_ECOLI	0	209	0	497	240387	238953
HI0342	NAPF_ECOLI	4.4 e −18	78	0	142	369091	369512
HI0498	POT2_HAEIN	0	312	0	353	514241	513183
HI0537	UREF_BACSB	1.2 e −22	194	3.3 e −31	227	561774	561094
HI0603	HEMX_ECOLI	3.3 e −15	219	0	230	632238	631166
HI0635	Y712_HAEIN	0	962	0	1093	677288	674105
HI0686	GLPT_ECOLI	7.2 e −15	43	0	474	729481	730900
H10723	TRKH_ECOLI	0	401	0	484	768811	770258
H10962	SYI_HAEIN	1 e −14	29	0	940	1021082	1018257
HI0976	YWFM_BACSU	1.9	94	1.2 e −11	291	1034007	1034833
HI1042	METH_ECOLI	5.6 e −25	144	0	591	1107721	1105899
HI1318	IF3_HAEIN	0	135	0	172	1394463	1394975
HI1383	PSTS_HAEIN	0	258	0	332	1479488	1478493
HI1390	YFRC_PROVU	32	16	2.2 e −21	87	1487105	1487382
HI1460	YADA_YEREN	—[a]	—[a]	2.8 e −06	173	1542922	1542416
HI1537	LICA_HAEIN	0	267	0	319	1608029	1608985
HI1653	TLDD_HAEIN	0	122	0	483	1719322	1717878
HI1721	YI5B_ECOLI	0	216	0	276	1793051	1793866
Methanococcus jannaschii							
MJ0165	PUR6_METTH	1.3 e −05	132	2.7 e −06	174	169019	168513
MJ0910	BCHD_RHOCA	1.8 e −07	133	1.8 e −14	214	842178	841567
MJ1209	MTH1_HAEPA	5.6 e −08	92	1.4 e −17	231	1153974	1153292
MJ1268	GLTL_ECOLI	3.2 e −10	182	1.5 e −20	231	1212681	1213378
MJ1325	CADF_STAAU	8.9 e −13	77	1.5 e −13	108	1275358	1275681
MJ1328	MTHC_HAEIN	9.7 e −12	214	0	485	1277831	1279247
MJ1339	IF2_BACST	2.6 e −09	117	1.1 e−08	165	1287412	1287870
MJ1353	FDHA_METFO	0	541	0	680	1304037	1302022
Mycoplasma genitalium							
MG434	PYRH_ECOLI	9.9 e −18	102	1.2 e−30	221	540530	541189

Note. Summary of FASTX search results using extended ORFs. All FASTX searches were done against SwissProt Release 34.

[a] With HI1460 ORF as a query, YADA_YEREN is not detected as a match with $E() < 100$.

sequences can suffer from a number of potential errors. Perhaps the most obvious source of error is mistakes in the sequence itself. Specifically, deletions and insertions resulting in frameshift errors can result in incorrect ORF assignment or in failure to identify the correct homologue. Another source of error in genomic analysis is the failure to identify the correct boundaries of an ORF. This may occur, for example, when the computer has a choice between two different start codons and chooses the wrong one. A third source of error is the failure to identify an ORF entirely. All of these errors can potentially be remedied by searching different combinations of genomic sequence with FASTX. The obvious limitation of these methods is that FASTX can only identify errors in sequences that have homologues in the protein databases. In the case of *M. jannaschii,* this may be a limiting factor, since less than 50% of its ORFs have identifiable homologues (Bult *et al.*, 1996; Kyrpides *et al.*, 1996).

Preliminary results of potential ORF boundary extensions are shown in Table 1. In most cases, it is difficult to determine from the FASTX alignments alone whether detected frameshifts and internal stop codons are the result of sequencing errors or are actual mutations in the original sequence. For instance, MJ1209 matches a methylase from *Haemophilus parainfluenzae* (MTH1_HAEPA) with an expectation value of 1.4×10^{-17}; however, the alignment is 83 amino acid residues short of the full length of MTH1_HAEPA and contains five internal stop codons. This extended ORF may well be a pseudogene. In other instances, the detected extensions are almost certainly the result of sequencing or record-keeping errors. For example, HI1653 matches TLDD_HAEIN with 100% identity over the full length of the match except for one frameshift. These extensions can also aid in revealing previously unknown homologues. HI0117 is currently described as “No database match” in the on-line documentation at the TIGR

TABLE 2

New Genes between Open Reading Frames

Name	Match	Fragment $E()$	Fragment Length	Start	Stop
Haemophilus influenzae					
HI1462.1	YD38_HAEIN	0	164	1546020	1545529
Methanococcus jannaschii					
MJ0469.1	RS14_METVA	7.8 e −19	53	415655	415813
MJ1158.1	YACA_BACSU	1.5 e −08	233	1097784	1098194
MJ1176.1	CAPM_STAAU	7.2 e −14	347	1117375	1116449
MJ1178.1	YCT2_BACFI	1.6 e −12	232	1117849	1118436
MJ1188.1	DRAG_RHORU	1.2 e −20	287	1127135	1127980
Mycoplasma genitalium					
MG291.1	YQGF_HAEIN	0.00029	136	357877	357486
MG335.1	Y060_MYCGE	0.00025	103	420345	420641
MG334.2	YOXG_BACSU	4.7 e −07	59	420235	420059

Note. FASTX searches that revealed previously unknown genes. Fragment *E*() values refer to the entire intergenic fragment, not just the alignment region given. "Length" corresponds to the length of the alignment in amino acid residues.

WWW site, but here we show that if extended downstream, HI0117 matches a membrane-bound lytic murein transglycosylase A precursor from *Escherichia coli* (MLTA_ECOLI) with an expectation value of 0.

Potential genes not previously identified are summarized in Table 2. Only the clearest examples are shown here; more than 60 additional intragenic regions contain translation frames with significant similarity to an entry in SwissProt, but the basis for the similarity was less clearcut (typically the sequence similarity was much shorter than expected). Identification of potential start and stop codons for these ORFs has not been completed at this time; thus, the start and stop positions listed are for alignments only. For consistency, the naming of the new genes follows the naming convention of TIGR; for example, the gene found between HI1462 and HI1463 is here named HI1462.1. This particular ORF is 100% identical to the DNA sequence of HI1338 and yet remained unidentified both in the initial analysis (Fleischmann *et al.*, 1995) and in the analysis of Robison *et al.* (1996). In the *M. jannaschii* genome, a ribosomal protein S14 homologue (MJ0469.1) has been identified among a large cluster of other ribosomal proteins (MJ0465–MJ0477). Note, also, that the results in Tables 1 and 2 are based on searches of SwissProt. Searches of the PIR or GenPept databases may reveal additional ORFs or ORF extensions.

DISCUSSION

FASTX, FASTY, TFASTX, and TFASTY can provide a sensitive tool for rapidly characterizing DNA sequence similarities based on translated DNA–protein sequence comparisons. Translated DNA–protein sequence comparisons are almost as sensitive as protein database searches (Figs. 4, 8) and dramatically more sensitive than DNA database searches. None of the relationships in Fig. 4 based on ancestors shared >500 Myr ago could be detected by DNA sequence comparison.

FASTX and FASTY provide two slightly different strategies for aligning the codons of a DNA sequence with a protein sequence. In our tests, FASTY is about 27% slower than FASTX. FASTX and FASTY show very similar performance tests with the error rates shown here. In other tests with much higher deletion rates (5–20% deletion only, rather than the 0.35%–3.5% deletions shown here), FASTY was significantly better (data not shown). However, in all of our tests with mixtures of insertions, deletions, and substitutions, there was little difference in performance between FASTX and FASTY. Thus, FASTX is preferred for its speed, although FASTY is capable of producing more accurate alignments.

Unlike conventional protein similarity searches, which are only moderately affected by changes in gap-penalty values (Pearson, 1995), the best frameshift penalties are dependent on the expected error level. Thus, for error-free data, a high frameshift penalty is appropriate; but if there is 2–5% error, lower frameshift penalties are more effective. For general purpose searching on high quality databases, such as the genomic scans summarized here, a high frameshift penalty (−25 to −30) is effective. For EST searches, where errors are expected, a lower penalty (−15 to −20) may be more appropriate.

Accurate estimates of statistical significance are essential for automatic large-scale sequence identification. FASTX and FASTY can provide accurate estimates if they are used to search databases from which low-complexity regions have been removed (Fig. 8). Likewise, TFASTX can produce reliable estimates if the protein query sequence does not contain low-complexity regions. Future modifications to FASTX/Y and TFASTX/Y may include a **seg**-style screening of the sequences used in the initial lookup table, to attempt to avoid detection of out-of-frame low-complexity runs in the initial comparison phase. This may make it possible to search unmodified protein databases with reliable statistics.

We were surprised to learn that so many gene-boundary assignments in the recently determined bacterial genomes were suspect, and that additional genes with very significant similarity to known proteins could be found in intergenic regions. If only 50–70% of the genes in *M. jannaschii* have clear homologues in the protein databases, then we may have detected only about 1/2 of the errors. We suspect that many of these errors occurred because a full Smith–Waterman DNA–protein alignment was not performed on initial gene assignments. Even if FASTX is not used initially to characterize open reading frames, it should be used in a final step to test for potential extensions of the translated DNA alignment.

ACKNOWLEDGMENTS

This work was supported by grants from the National Library of Medicine (LM04969) to W.R.P. and (LM05110) to W.M.

REFERENCES

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sulton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reisch, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273:** 1058–1073.

Chao, K.-M., Pearson, W. R., and Miller, W. (1992). Aligning two sequences within a specified diagonal band. *Comp. Appl. Biosci.* **8:** 481–487.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., amd M. D. Cotton, E. H., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., III, C. A. H., and Venter, J. C. (1995). The minimal gene complement of Mycoplasma genitalium. *Science* **270:** 397–403.

Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genet.* **3:** 266–272.

Guan, X., and Uberbacher, E. (1996). Alignments of DNA and protein sequences containing frameshift errors. *Comp. Appl. Biosci.* **12:** 31–40.

Hein, J. (1994). An algorithm combining DNA and protein alignment. *J. Theor. Biol.* **167:** 169–174.

Hein, J., and Stovbaek, J. (1994). Genomic alignment. *J. Mol. Evol.* **38:** 310–316.

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

Hillier, L., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Soares, M. B., Tan, F., Thierry-Meg, J., Trevaskis, E., Underwood, K., Wohldman, P., Waterston, R., Wilson, R., and Marra, M. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Huang, X., and Zhang, J. (1996). Methods for comparing a DNA sequence with a protein sequence. *Comp. Appl. Biosci.* **12:** 497–506.

Knecht, L. (1995). Pairwise alignment with scoring on tuples. *In* "Lecture Notes in Computer Science: Combinatorial Pattern Matching," Vol. 937, pp. 215–229, Springer-Verlag, Berlin.

Koonin, E. V., Mushegian, A. R., Tatusov, R. L., Altschul, S. F., Bryant, S. H., Bork, P., and Valencia, A. (1994). Eukaryotic translation elongation factor 1 gamma contains a glutathione transferase domain—Study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.* **3:** 2045–2054.

Kyrpides, N. C., Olsen, G. J., Klenk, H.-P., White, O., and Woese, C. R. (1996). Methanococcus jannaschii genome: Revisited. *Microb. Comp. Genomics* **1:** 329–338.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *In* "Methods in Enzymology" (R. F. Doolittle, Ed.), Vol. 183, pp. 63–98, Academic Press San Diego.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4:** 1145–1160.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266:** 227–258.

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

Peltola, H., Soderlund, H., and Ukkonen, E. (1986). Algorithms for the search of amino acid patterns in nucleic acid sequences. *Nucleic Acids Res.* **14:** 99–107.

Robison, K., Gilbert, W., and Church, G. M. (1996). More *Haemophilus* and *Mycoplasma* genes. *Science* **271:** 1302–1303.

Wootton, J. C. (1994). Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.

Zhang, Z., Pearson, W., and Miller, W. (1997). Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.* **4:** 337–343.