[5] Rapid and Sensitive Sequence Comparison with FASTP and FASTA

By WILLIAM R. PEARSON

Introduction

Rapid computer algorithms for comparing DNA and protein sequences have dramatically decreased the amount of time required to compare an unidentified sequence to a DNA or protein sequence database. These methods have come at a time when the protein and DNA sequence libraries are growing almost 50% per year, owing to more efficient cloning techniques and more productive sequencing procedures. Sequence searches have led to the discovery of many new families of proteins, including the tyrosine kinase oncogene family, the steroid receptor and v-*erbA* oncogene family, the growth factor receptor family, the G-protein-coupled receptor family, and transcription factors containing a zinc-finger motif. Often a database search provides the first insight into the mechanism of action of a newly sequenced protein.

In 1985, David Lipman and I described the FASTP program for searching protein sequence libraries.[1] FASTP combines a rapid technique for focusing on those regions in a pair of sequences that share a high density of identities with a scoring procedure that uses the PAM250 scoring matrix[2] (Fig. 3) for high sensitivity. FASTP decreased the computer time required for a protein database search from about 6 h on a VAX11/780 to about 10 min on an IBM-PC. Shortly thereafter, I modified FASTP for DNA sequence comparison and distributed the FASTN program. More recently, we described an improved version of FASTP called FASTA, which combines the functions of the FASTP and FASTN programs and provides a more sensitive sequence comparison algorithm.[3] In addition, the FASTA package (Table I) includes programs for comparing a protein sequence to a DNA sequence database (TFASTA), for identifying local sequence similarities or duplications in sequences (LFASTA, PLFASTA), and for evaluating the statistical significance of a similarity score (RDF2). The FASTA programs can be tailored to specific comparison problems by

[1] D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985).

[2] M. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *in* "Atlas of Protein Sequence and Structure" (M. Dayhoff, ed.), Vol. 5, Suppl. 3, p. 345. National Biomedical Research Foundation. Silver Spring, Maryland, 1978.

[3] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (1988).

TABLE I
FASTA SEQUENCE COMPARISON PROGRAMS

Program	Description
FASTA	Compares a protein sequence to another protein sequence or to a protein database, or a DNA sequence to another DNA sequence or to a DNA library.
TFASTA	Compares a protein sequence to a DNA sequence or DNA sequence library. The DNA sequence is translated in all six reading frames, and the protein query sequence is compared to each of the six derived protein sequences. The DNA sequence is translated from one end to the other; no attempt is made to edit out intervening sequences. Termination codons are translated into unknown ("X") amino acids. The eukaryotic nuclear genetic code is used for all translations.
LFASTA	Compares two sequences to identify regions of sequence similarity. While FASTA and TFASTA report a single alignment between two sequences, LFASTA will report several sequence alignments if there are several similar regions. LFASTA can identify similarities arising from internal repeats or similar regions that cannot be aligned by FASTA because of gaps (Fig. 6). LFASTA reports actual sequence alignments and similarity scores.
PLFASTA	PLFASTA is identical to LFASTA, but it presents a dot-matrix-like plot of the similar regions, rather than the actual alignments. PLFASTA plots are shown in Figs. 6 and 7.
RDF2	Evaluates the significance of pairwise similarity scores using a Monte Carlo analysis. Similarity scores for the two sequences are calculated, and then the second sequence is shuffled 50 to 200 times and compared with the first sequence. RDF2 can use one of two shuffling strategies. One strategy simply keeps the amino acid composition of the entire shuffled sequence identical to the unshuffled sequence. The second local shuffle destroys the order but preserves the composition of small segments (10–25 residues) of the shuffled sequence.

changing the similarity scoring matrix and gap penalties. Thus, the same program can be used to compare protein sequences with the PAM250 matrix or a matrix based on the genetic code and to compare DNA sequences. In addition, FASTA provides several output options that can be used to highlight similarities and differences in aligned sequences.

In this chapter, I show an example of a simple FASTA library search, describe the FASTA algorithm, and then discuss in detail a more problematic search, namely, one for members of the G-protein-coupled receptor family. Additional information about how to customize the scoring parameters and output from the FASTA programs is included in the appendices.

Much of this chapter focuses on the evaluation of distant alignments that have ambiguous similarity scores. In any database search, there is always a library sequence with the best score, regardless of whether that sequence shares common ancestry or some other significant similarity with the query sequence. In sequence comparison, there is a trade-off between

sensitivity—the ability to identify distantly related sequences—and selectivity—the avoidance of false positives (unrelated sequences with high similarity scores). The perfect sequence comparison method would be both sensitive and selective; it would rank all the members of a protein family that share a common ancestor above all the sequences that are similar but nonhomologous. No such program exists, because proteins evolve at very different and sometimes very rapid rates, so that a sequence may contain only a trace of its evolutionary ancestry. In many cases, the problem is to differentiate between high scoring sequences that share common ancestry or significant similarity with the query sequence and sequences with high scores that are due to local sequence composition and random chance. The FASTA program is more sensitive than FASTP, so more distantly related sequences can be identified with FASTA. Nevertheless, methods that increase sensitivity decrease selectivity, and additional care is required when interpreting the results of a FASTA search.

Throughout this chapter, the emphasis is on protein sequence comparison. FASTA can compare either DNA or protein sequences, but protein sequence comparison is far more useful, because distant sequence relationships can best be identified at the protein sequence level. While DNA sequences that encode structural RNAs have been successfully used to examine ancient evolutionary relationships, DNA sequences from repeated sequences, intervening sequences, upstream regions, or untranslated regions of messages rarely allow the demonstration of common ancestry for sequences that diverged more than 100 to 200 million years ago. In contrast, common ancestry can frequently be demonstrated for protein sequences that diverged 1 to 2 billion years ago. Protein sequence comparisons are more useful both because of the degeneracy of the genetic code (a change in a DNA sequence may not change the encoded protein) and also because of the biochemical information in the amino acid itself (arginines are very similar to lysines, but glycines and isoleucines are very different). If there is ever a question about the relationship between two DNA sequences which encode proteins, the comparison should always be done with the derived amino acid sequences.

Using FASTA Programs

Although the FASTA programs provide a number of options for customizing searches, most of the time only three entries are required: the name of the file containing the query sequence, the name of the file containing the library or a second sequence, and the value of the *ktup* parameter. FASTA and TFASTA compare the first (query) sequence to all the sequences in the second file (there need only be one), reporting the one best

similarity score and alignment for each pairwise comparison. All of the FASTA programs calculate a "local" similarity score, i.e., the best region of similarity is found between the two sequences being compared. The score of the local region is not affected by poorly aligned portions of the sequences outside the best region. Thus, programs in the FASTA package can be used to find conserved or shuffled protein domains, such as the epidermal growth factor (EGF) precursorlike domains in the low density lipoprotein (LDL) receptor[4] (Fig. 7).

FASTA and TFASTA report only the similarity score for the one best pairwise alignment between two sequences. In the case of proteins with repeated domains, there may be several alignments with high similarity scores that are of biological interest. Multiple regions of similarity can be displayed as alignments or as a dot-matrix-style plot by LFASTA and PLFASTA, respectively. LFASTA and PLFASTA use a slight modification of the FASTA algorithm that focuses more tightly on local regions of similarity, so that regions of strong similarity do not overshadow neighboring regions with lower similarity scores.

RDF2 is designed to evaluate the statistical significance of a pairwise similarity score. The program calculates similarity scores for the best pairwise alignment, using the FASTA algorithm, then randomly shuffles the second sequence and calculates pairwise scores for the query sequence and each of the shuffled sequences. By examining the distribution of similarity scores obtained with randomly shuffled sequences with the same length and amino acid composition, one can evaluate the likelihood that the similarity scores for the unshuffled sequence are due to unusual sequence composition or other random fluctuations.

Figure 1 shows an example of a search of the National Biomedical Research Foundation Protein Identification Resource (PIR) library, using a murine glutathione transferase as the query sequence. After the program is started, it asks for three entries: the query sequence file name (gst87.aa in Fig. 1), selection of the sequence library (p for the NBRF protein database), and the *ktup* parameter (2 in this case by default). In this example, the user is prompted for the file names and *ktup* parameter, but each of these entries can be specified on the command line. For example, the command

```
fasta gst87.aa p 1
```

would do the same search with $ktup = 1$.

After the search is finished, the program asks for a file name for the results (gt875.k2 in Fig. 1). If a file name is specified, the histogram, the list of top scoring library sequences, and the sequence alignments are written

[4] T. C. Sudhof, J. L. Goldstein, M. S. Brown, and D. W. Russell, *Science* **228**, 815 (1985).

to the file as well as to the display terminal. If the results are written to a file, the sequence alignments are not shown on the terminal. The program then asks for the number of sequences to be displayed initially. In this example, there were 2634 library sequences with initial similarity scores greater than 28, and as many as 2634 scores could be displayed.[5] The meanings of the terms initial similarity score *(initn), init1* score, and optimized score are discussed in the next section.

The example in Fig. 1 is a simple one; the 20 top scoring sequences have similarity scores that indicate common ancestry (homology), and the rest of the sequences are unrelated. Mammalian glutathione transferases can be grouped into three classes; members of the same class share 80–95% amino acid sequence identity, while interclass alignments show 25–30% sequence identity.[6] Members of all three glutathione transferase classes are found in the list of top scoring sequences. Many searches of the protein sequence library, particularly with soluble proteins, are similar to this one, i.e., related library sequences have similarity scores that are well separated from the main distribution of similarity scores for unrelated sequences. When this occurs, one can be confident that a homologous sequence has been found. Alternatively, as was the case when this search was done in 1985, there may be no sequences with similarity scores greater than 60 because there are no related sequences in the library. Unfortunately, there are sometimes related sequences with similarity scores of 60 or less, and unrelated sequences with scores greater than 100.

FASTA Implementation

The programs are written in the C programming language and run on IBM-PC microcomputers under the DOS operating system, on the Macintosh, on computers running the System V, BSD, and Xenix versions of the Unix operating system, and on VAX computers running the VMS or Unix operating systems. VMS versions of the program are designed to be used in conjunction with the National Biomedical Research Foundation Protein Identification Resource PSQ and NAQ programs and with the University of Wisconsin Genetics Computer Group program package. The program code is very portable; exactly the same source code compiles on all the

[5] This search was done on a Sun 3/260 workstation. On an IBM-PC, no more than 2000 scores would be saved. The older FASTP program would save the first 1000–2000 scores greater than the CUTOFF value and then fail to save additional high scoring sequences. The FASTA program saves the first 6000 (2000 on an IBM-PC) scores greater than the CUTOFF value, and, if additional high scoring sequences are found, it saves the top 75% of the sequences and adjusts the CUTOFF value upward. Thus, FASTA always saves the top scoring library sequences; in some cases FASTP does not.

[6] B. Mannervik and U. H. Danielson, *CRC Crit. Rev. Biochem.* **23**, 283 (1988).

```
Visage 2000 % fasta
 fasta 1.3 [Feb, 1989] searches a sequence data bank

Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

 test sequence file name: gst87.aa

 Choose sequence library:

    G: GENBANK Translated Protein Database
    P: NBRF Protein Database
    A: NBRF Protein Database + Genbank
    S: Swiss-Prot Release 8

 Enter letter or filename: p
 ktup? (1 to 2) [2] __
 >GT8.7 transl. of pa875.con, 19 to 675 : 217 aa      sequence description
 vs NBRF Protein Database library                     Library description
 searching aabank.19 library                          Library file name
  2802054 aa in 10526 sequences                       Search is finished

      initn   init1
<  2     4     4:==
   4     0     0:
   6     6     6:===
   8    16    16:========
  10    42    42:====================
 -----//-----
  40   140   124:==================================================
  42   115    72:------------------------------------+++++++++++++
  44    86    42:--------------------+++++++++++++++++++++
  46    63    29:---------------+++++++++++++++++
  48    59    15:--------++++++++++++++++++++++
  50    36    10:-----++++++++++++++
  52    21     9:-----++++++                    '-' indicates the number of init1 scores
  54    13     2:-++++++
  56     9     5:---++
  58    16     2:-+++++++                       '+' indicates the number of initn scores
  60     7     0:++++
  62     2     0:+
  64     7     0:++++
  66     6     0:+++
  68     1     0:+
  70     0     0:
  72     0     0:
  74     3     0:++
  76     0     0:
  78     0     0:
  80     0     1:-
> 80    20    19:==========
2802054 residues in 10526 sequences
 mean initn score:  25.2 (7.08)
 mean init1 score:  24.9 (6.38)
 2634 scores better than 28 saved, ktup: 2, fact: 8  scan time:  0:01:14
 Enter filename for results : gt875.k2
 How many scores would you like to see? [20] 25
```

FIG. 1. Sample FASTA library search. A transcript of a protein sequence library search with the FASTA program is shown. Entries typed by the user are underlined. A large portion of the histogram of library similarity scores has been removed; complete histograms are shown in Fig. 4. The alignments are written to a file in this example. The selection of the library to be searched from a list of libraries is made possible by defining the FASTLIBS environment symbol.

```
The best scores are:                                      initn initl   opt
A25510 - Glutathione S-transferase (EC 2.5.1.18), Yb c 1133  1133  1133
A24085 - Glutathione S-transferase (EC 2.5.1.18), Yb1  1127  1127  1127
B26187 - Glutathione S-transferase (EC 2.5.1.18), Yb-2 1018  1018  1038
A26307 - Glutathione S-transferase (EC 2.5.1.18), Yb2  1018  1018  1038
XURTG4 - Glutathione S-transferase (EC 2.5.1.18), chai  999   999  1019
A29036 - Glutathione S-transferase (EC 2.5.1.18) Yb3 -  988   988  1010
A26484 - Glutathione S-transferase (EC 2.5.1.18) - Flu  503   416   545
XURTGP - Glutathione S-transferase P (EC 2.5.1.18) - R  198   109   289
B20831 - Glutathione S-transferase (EC 2.5.1.18) minor  189   189   200
A20831 - Glutathione S-transferase (EC 2.5.1.18) major  174   174   180
H24735 - Glutathione transferase (EC 2.5.1.18), MIII -  137   137   137
A28562 - Glutathione S-transferase (EC 2.5.1.18), clas  134   134   144
E24735 - Glutathione transferase (EC 2.5.1.18), class   130   130   130
B22457 - Glutathione S-transferase (EC 2.5.1.18) mu -   130   130   130
J24735 - Glutathione transferase (EC 2.5.1.18), GT-9.3  126   126   126
G24735 - Glutathione transferase (EC 2.5.1.18), 4-4 -   125   125   126
I24735 - Glutathione transferase (EC 2.5.1.18), GT-8.7  116   116   120
K24735 - Glutathione transferase (EC 2.5.1.18) - Bovin  107   107   107
F24735 - Glutathione transferase (EC 2.5.1.18), 3-3 -   104   104   114
A26598 - 28K antigen precursor - Fluke (Schistosoma ma   87    79   131
A29944 - Chaoptin precursor - Fruit fly                  74    51    82
A29352 - SST2 protein - Yeast (Saccharomyces cerevisia   73    52    54
QQECO3 - Hypothetical protein F-300 - Escherichia coli   73    45    59
A29949 - Glycogen phosphorylase (EC 2.4.1.1), brain -    68    52    70
A25026 - Chloramphenicol acetyltransferase (EC 2.3.1.2   66    43    47
 More scores? [0] __
 Display alignments also? y
 number of alignments [20]? 20
Library scan:  0:01:14  total CPU time:  0:01:20
```

FIG. 1.

machines except the Macintosh. Versions of the program for the IBM-PC or Macintosh can search the library with query sequences up to 2,000 residues in length and can search library sequences of any length. (For the IBM-PC, this limit is set so that the programs can run using the faster small memory model, which limits data to 64 kilobytes). On other machines, the query sequence can be up to 10,000 residues, but this value can be increased by editing and recompiling the programs. There is no limit to the length of the library sequence on large or small machines. If the library sequence is too long (>10,000 residues for small machines, >50,000 for large), it is scanned in overlapping pieces. The complete source code is available for all versions of the program from William R. Pearson.

The searching programs FASTA and TFASTA can search libraries in a variety of different formats, including: (1) FASTP/DM or query sequence format; (2) GenBank magnetic tape format; (3) the Protein Identification Resource CODATA format; (4) EMBL and SWISS-PROT format; (5) IntelliGenetics sequence file format; and (6) Compressed GenBank format for floppy disk distribution (Appendix 1). Although the programs can be

run immediately after they are copied onto a computer, they are easier to use if an additional file is installed. This file, which is referred to by the UNIX, DOS, or VMS environment symbol FASTLIBS, allows FASTA and TFASTA to list the libraries that are available to be searched. FASTA can also search a library made up of several files containing data, as is often the case for libraries that are distributed on floppy diskettes. To indicate that a library file contains a list of file names, rather than actual sequence data, the library file name is preceded by the symbol @.

The behavior of all the FASTA programs can be modified by specifying a different scoring matrix file or CUTOFF value, and the alignments displayed by FASTA, TFASTA, and LFASTA can be modified by specifying options on the command line or with environment symbols. A complete list of input and output options is described in Appendix 3.

FASTA Algorithm

FASTA uses four steps to calculate three scores that characterize sequence similarity; these steps are outlined in Table II. The first step uses a rapid technique for finding identities shared between two sequences; the method is similar to an earlier technique described by Wilbur and Lipman.[7] FASTP and FASTA achieve much of their speed and selectivity in this first step by using a lookup table[8] to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison.[9] The *ktup* parameter determines how many consecutive identities are required in a match. A *ktup* value of 2 is frequently used for protein sequence comparison, which means that the program examines only those portions of the two sequences being compared that have at least two adjacent identical residues in both sequences. More sensitive searches can be done using $ktup = 1$. For DNA sequence comparisons, the *ktup* parameter can range from 1 to 6; values between 4 and 6 are recommended. When the query sequence is a short oligonucleotide or oligopep-

[7] W. J. Wilbur and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726 (1983).

[8] A lookup table is a rapid method for finding the position of a residue in a sequence. One way to find the "A" in the sequence "NDAPL" is to compare the "A" to each residue in the sequence. A faster way (if many residues are to be checked), is to make a table of all possible residues (23 for proteins) so that the computer representation for the residue (e.g., "A" is 1, "R" is 2, "N" is 3) is the same as its position in the table. A value is then placed in the table that indicates whether the residue is present in the sequence and, if it is, where it is present. For this example, the table has the value 1 at position 3 in the table ("N" is the third amino acid), 2 at position 4, 3 at position 1, 4 at 15, 5 at 11, and the remaining 18 positions are 0. The presence and position of the "A" in the sequence can then be determined in a single step by looking it up at position 1 in the table.

[9] J. P. Dumas and J. Ninio, *Nucleic Acids Res.* **10**, 197 (1982).

TABLE II
CHARACTERIZATION OF SEQUENCE SIMILARITY BY FASTA

Step 1	Identify regions shared by the two sequences with the highest density of identities (*ktup* = 1) or pairs of identities (*ktup* = 2).
Step 2	Rescan the ten regions with the highest density of identities using the PAM250 matrix shown in Fig. 3. Trim the ends of the region to include only those residues contributing to the highest score. Each region is a partial alignment without gaps.
Step 3	(FASTA only) If there are several initial regions with scores greater than the CUTOFF value, check to see whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined initial regions minus a penalty (usually 20) for each gap. This initial similarity score (*intin*) is used to rank the library sequences. The score of the single best initial region found in Step 2 is reported (*init1*); it is the same as the initial similarity score calculated by FASTP.
Step 4	Construct a NWS optimal alignment of the query sequence and the library sequence, considering only those residues that lie in a band 32 residues wide centered on the best initial region found in Step 2. FASTA and FASTP both report this score as the optimized (*opt*) score.

tide, *ktup* = 1 should be used. A sequence comparison using the traditional Needleman–Wunsch–Sellers (NWS) algorithm[10,11] requires a number of residue comparisons proportional to the product of the lengths of the sequences being compared, for example, 33,434 for a comparison of hemoglobin β chain (146 amino acids) with trypsin (229 amino acids). (Methods that compare all the fixed length segments of one sequence with another require a similar amount of time.) FASTP and FASTA require only 94 comparisons (1/355th as many) to examine β-globin and trypsin with *ktup* = 2; this number increases to 1921 with *ktup* = 1.[12]

In conjunction with the lookup table, we use the "diagonal" method to find all regions of similarity between the two sequences, counting *ktup* matches and penalizing for intervening mismatches.[1,7] This method identified regions of a diagonal that have the highest density of *ktup* matches. The term diagonal refers to the diagonal line that is seen on a dot matrix plot[13] when a sequence is compared with itself, and it denotes an alignment between two sequences without gaps (see Figs. 6 and 7). For example, the alignment of residue 8 in sequence one (on the x axis) and 13 in sequence

[10] S. Needleman and C. Wunsch, *J. Mol. Biol.* **48**, 444 (1970).

[11] P. Sellers, *SIAM J. Appl. Math.* **26**, 787 (1974).

[12] The ratio of comparisons with *ktup* = 2 to *ktup* = 1 is 20, exactly the value predicted by the number of different amino acids. Additional processing is required for each *ktup* = 2 match, however, and in practice the time required for protein searches with *ktup* = 1 is only about 5 times that with *ktup* = 2.

[13] J. Maizel and R. Lenk, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7665 (1981).

two (y axis) specified a diagonal; the alignment of residue 16 (sequence one) and 21 (sequence two) falls on the same diagonal, as do 27 and 32, and so on.

As shown in Fig. 2, FASTP uses a simple formula to identify portions of a diagonal with a high density of identities, referred to as a local region of

For each *ktup* group of residues in the library sequence {

A. For each position in the query sequence with the identical residues {

1. Calculate the diagonal based on the current position in the library and query sequences; (diagonal = library_position - query_position)
2. Check to see if an initial region has been found in this diagonal;
3. If no region has been found on this diagonal before, save the current query_position as the start of a region; the score for the initial region is 16 (for *ktup* =2 in FASTP).
4. Otherwise {
   a. If there already was an region on this diagonal, calculate the distance from the end of the previous initial region to the current *ktup* match.
   b. If that distance is greater than the score of the previous region, save the previous region and start a new region.
   c. Otherwise {
      i. Save the score of the previous region;
      ii. Extend the previous region to the current *ktup* match;
      iii. Decrease the score of the previous region by the distance the region was extended and increase it by 16 for the new *ktup* match.
      }
   }
}
}

Each time the program saves a region, it determines whether the score of the region is better than the lowest score of the ten best regions that have already been saved. If new score is better, it replaces the lowest scoring saved region with the new region, and finds new lowest scoring region in the updated list of regions.

FIG. 2. FASTP/FASTA scanning algorithm. The logic used to scan sequences in the library and identify regions of similarity is shown. The values shown for increasing the score of a region at a *ktup* match are used in FASTP. In FASTA, the values used to increment the score of a region at a *ktup* match are based on the PAM250 matrix. Thus, matching a Leu-Leu in the query sequence with a Leu-Leu in the library sequence would increase the region score by 24 instead of 16, while matching an Ala-Ser with an Ala-Ser would increase the region score by 8 instead of 16. As a result, the score of the initial region includes the PAM250 scores of all the *ktup* identities, with a constant penalty (−1) for each residue that separates *ktup* identities on the same diagonal. For DNA sequences, a constant value is used, which is equal to the square of the *ktup* value.

similarity, or simply a region. As a result of this formula, a group of *ktup* = 2 amino acid matches separated by fewer than 16 residues would be combined into a region, but *ktup* = 2 matches separated by more than 30 residues from the previous region would start a new region. FASTA uses a formula for scoring *ktup* matches that incorporates the actual PAM250 values for the aligned residues. Thus, groups of identities with high similarity scores contribute more to the local diagonal score than do identities with low similarity scores. This more sensitive formula is used for protein sequence comparisons; the constant value for *ktup* matches is used for DNA sequence comparisons. The earlier NUCALN programs[7] use a constant value for *ktup* matches, but these programs save the one best local region in each diagonal. FASTA saves the ten best local regions, regardless of whether they are on the same or different diagonals.

After the ten best local regions are found in the first step, they are rescored using a scoring matrix that allows runs of identities shorter than *ktup* residues and conservative replacements to contribute to the similarity score. For protein sequences, this score is usually calculated using the PAM250 matrix[2] (Fig. 3), although scoring matrices based on the minimum number of base changes required for a specific replacement, on identities alone, or on an alternative measure of similarity, can also be used with FASTA. The PAM250 scoring matrix was derived from the analysis of the amino acid replacements occurring among related proteins, and it specifies a range of positive scores for replacements that commonly occur among related proteins and negative scores for unlikely replacements. FASTA can also be used for DNA sequence comparisons, and matrices can be constructed that allow separate penalties for transitions and transversions. For each of the best diagonal regions rescanned with the scoring matrix, a subregion with the maximal score is identified.

The FASTP program uses the single best scoring initial region to characterize pair-wise similarity; the initial scores are used to rank the library sequences. The FASTP initial score is also calculated by FASTA, and it is referred to as the *init1* score. FASTA goes one step further during a library search; it checks to see whether several initial regions can be joined together in a single alignment to increase the initial score. Thus, FASTA improves on the sensitivity of FASTP by allowing multiple high scoring initial regions to be joined. Given the locations of the initial regions, their respective scores, and a "joining" penalty (analogous to a gap penalty), FASTA calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. This optimal alignment of initial regions can be rapidly calculated using a dynamic programming algorithm similar to that described for NUCALN;[7] FASTA uses the resulting score, referred to as the *initn* score, to rank the library sequences.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	???
A	Ala	2																					
R	Arg	-2	6																				
N	Asn	0	0	2																			
D	Asp	0	-1	2	4																		
C	Cys	-2	-4	-4	-5	12																	
Q	Gln	0	1	1	2	-5	4																
E	Glu	0	-1	1	3	-5	2	4															
G	Gly	1	-3	0	1	-3	-1	0	5														
H	His	-1	2	2	1	-3	3	1	-2	6													
I	Ile	-1	-2	-2	-2	-2	-2	-2	-3	-2	5												
L	Leu	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6											
K	Lys	-1	3	1	0	-5	1	0	-2	0	-2	-3	5										
M	Met	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6									
F	Phe	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9								
P	Pro	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6							
S	Ser	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2						
T	Thr	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3					
W	Trp	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17				
Y	Tyr	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10			
V	Val	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4		
B	Asx	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2	
Z	Glx	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3
X	???	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG.3. PAM250 scoring matrix. The amino acid replacement matrix developed by Dayhoff *et al.* [M. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *in* "Atlas of Protein Sequence and Structure" (M. Dayhoff, ed.), Vol. 5, Suppl. 3, p. 345. National Biomedical Research Foundation, Silver Spring, Maryland, 1978] and used by FASTP and FASTA is shown. The numbers indicate the score for a match between the amino acids in the same row and column. Tryptophan and cysteine receive the highest scores when paired with themselves in an alignment (17 and 12, respectively); the tryptophan/cysteine substitution receives the lowest score (−8). The PAM250 matrix scores the valine/isoleucine substitution with the same value (4) as the valine/valine identity.

This third "joining" step in the computation of the initial score increases the sensitivity of the search method because it allows for insertions and deletions as well as conservative replacements. The modification does, however, decrease selectivity, as can be seen in Fig. 4. We limit the degradation of selectivity by including in the optimization step only those initial regions whose scores are above an empirically determined threshold.[14]

After a complete search of the library, FASTA plots the initial scores of each library sequence in a histogram, calculates the mean similarity score for the query sequence against each sequence in the library, and determines the standard deviation of the distribution of initial scores (Fig. 4). The initial scores are used to rank the library sequences, and, in the fourth and final step of the comparison, the highest scoring library sequences are aligned using a modification of the standard NWS optimization method[10,15] (Tables III and IV). The optimization employs the same scoring matrix used in determining the initial regions; the resulting optimized alignments are calculated for further analysis of potential relationships, and the optimized similarity score is reported. With the FASTP program, optimization frequently improved the similarity scores of related sequences by factors of two or three; these improvements can be seen by comparing the *init1* score with the optimized score in Tables III and IV. Because FASTA calculates an initial similarity score based on an optimization of initial regions during the library search, the initial score is much closer to the optimized score for many sequences. In fact, unlike FASTP, the FASTA method may yield initial scores that are higher than the corresponding optimized scores.

Searching with FASTA: G-Protein-Coupled Receptors

In the late 1980s, molecular cloning of rhodopsin, the β-adrenergic receptor, and the acetylcholine receptor demonstrated the existence of a large superfamily of membrane-bound receptors that interact with guanine nucleotide regulatory proteins[16] (G-protein-coupled receptors). A common structure with seven conserved transmembrane regions connected by variable length cytoplasmic and extracellular loops has been predicted for these

[14] FASTA joins an initial region only if its similarity score is greater than the CUTOFF value, a value that is approximately one standard deviation above the average score expected from unrelated sequences in the library. For a 200-residue query sequence and *ktup* = 2, this value is 28.

[15] T. Smith and M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).

[16] R. J. Lefkowitz and M. G. Caron, *J. Biol. Chem.* **263**, 4993 (1988).

A. *ktup* =2

```
      initn   init1
<  2     1      1:=
   4     0      0:
   6     1      1:=
   8     2      2:=
  10     5      5:===
  12    28     28:==============
  14    40     40:====================
  16   122    122:============================================================
  18   202    202:============================================================
  20   415    415:============================================================
  22   836    836:============================================================
  24  1079   1079:============================================================
  26   884    884:============================================================
  28  1067   1067:============================================================
  30  1021   1097:============================================================
  32   435    519:============================================================
  34   365    460:============================================================
  36   212    301:============================================================
  38   183    203:============================================================
  40   190    168:============================================================
  42   126    103:============================================================
  44   104     47:------------------------++++++++++++++++++++++++++
  46    74     50:------------------------++++++++++++
  48    85     39:--------------------++++++++++++++++++++++
  50    68     13:-------++++++++++++++++++++++++++++
  52    36     12:------++++++++++++
  54    38      2:-++++++++++++++++++
  56    15      2:-+++++++
  58    12      1:-++++++
  60    15      2:-+++++++
  62     8      5:---+
  64    10      1:-++++
  66     6      0:+++
  68     3      1:-+
  70     2      0:+
  72     6      1:-++
  74     5      0:+++
  76     2      0:+
  78     1      2:=
  80     1      0:+
  82     1      0:+
  84     0      0:
  86     1      0:+
  88     1      0:+
  90     0      0:
  92     0      0:
  94     1      0:+
  96     0      0:
  98     0      0:
 100     0      0:
>100    15     13:-------+
2224465 residues in  7724 sequences,  scan time:  0:03:07
 mean initn score:  27.8 (7.36), mean init1 score:  27.2 (6.14)
```

FIG. 4A. See legend on p. 78.

B. *ktup* =1

```
      initn   init1
< 2       0      0:
  4       0      0:
  6       2      2:=
  8       0      0:
 10       1      1:=
 12       3      3:==
 14       7      7:=====
 16       9      9:======
 18      28     28:==============
 20      44     44:======================
 22      92     92:==============================================
 24     191    191:==============================================
 26     337    337:==============================================
 28     513    513:==============================================
 30     785    785:==============================================
 32     773    773:==============================================
 34     854    865:==============================================
 36     712    820:==============================================
 38     620    806:==============================================
 40     449    644:==============================================
 42     385    591:==============================================
 44     245    387:==============================================
 46     142    264:==============================================
 48     135    198:==============================================
 50     144    110:==============================================
 52     121     64:--------------------------------++++++++++++++++++
 54     179     66:--------------------------------++++++++++++++++++
 56     148     30:-------------+++++++++++++++++++++++++++++++++++++
 58     140     22:-----------+++++++++++++++++++++++++++++++++++++++
 60     118     19:----------++++++++++++++++++++++++++++++++++++++++
 62      94     12:------++++++++++++++++++++++++++++++++++++++++
 64      69      9:-----+++++++++++++++++++++++++++++++
 66      57      4:--++++++++++++++++++++++++++++
 68      47      3:--++++++++++++++++++++++
 70      42      4:--++++++++++++++++++++
 72      50      1:-++++++++++++++++++++++++++
 74      28      0:+++++++++++++++
 76      26      1:-++++++++++++
 78      26      0:++++++++++++
 80      15      1:-+++++++
 82      17      0:+++++++++
 84      11      1:-+++++
 86      11      0:++++++
 88       6      0:+++
 90       6      1:-++
 92       3      0:++
 94       6      0:+++
 96       1      0:+
 98       3      0:++
100       2      0:+
>100     27     16:--------++++++
2224465 residues in  7724 sequences, scan time:  0:17:08
 mean initn score:  36.1 (8.91), mean init1 score:  34.6 (6.62)
```

FIG. 4B. See legend on p. 78.

proteins by analogy with bacteriorhodopsin. Figures 4 and 5 and Tables III and IV show the results of a search of the SWISS-PROT protein sequence database using the β_2-adrenergic receptor as the query sequence. Examples are shown with both *ktup* = 2 and *ktup* = 1 to highlight the trade-offs between sensitivity and selectivity encountered during library searches.

Figure 4 shows the distribution of initial similarity scores calculated by FASTA for searches of the protein database with *ktup* = 2 and *ktup* = 1. Two distributions of scores are plotted, the distribution of the FASTA initial similarity score used to rank the library sequences *(initn)*, and the distribution of the older FASTP initial score *(init1)*. For example, in Fig. 4A there were 15 library sequences with *initn* scores of 59 or 60, but there were only 2 library sequences with *init1* scores in this interval. The values in the *initn* and *init1* columns are identical in the intervals including scores of 28 and below, because these scores fall below a threshold value (29 for this query sequence with *ktup* = 2). This example begins to show the effect of increasing the sensitivity of the search. With *ktup* = 2, there are 63 library sequences with *initn* scores greater than 60, while there are only 23 library sequences with *init1* scores greater than 60. The FASTA joining has moved scores out of the intervals between 29 and 38 into the intervals 39 and above.

The top ranking library sequences from the search with *ktup* = 2 are shown in Table III. In this example, all 17 of the top scoring sequences are G-protein-coupled receptors; only 2 of the 20 sequences with the highest similarity scores do not belong to this receptor family. Nevertheless, there are additional G-protein-coupled receptors that are not found in the top 40 sequences. One *Drosophila* opsin is ranked 98th, and a second (OPS3$DROME) is not found in the top 200 sequences with *ktup* = 2.

FIG. 4. Identification of sequences related to the β-adrenergic receptor. The β-adrenergic receptor (PIR entry QRHYB2, SWISS-PROT entry B2AR$MESAU, 418 amino acids) was used to search the Swiss-Prot protein sequence database (Release 8, August 1988), using the PAM250 scoring matrix. A total of 2,224,465 residues in 7,724 sequences were compared in 3 min (*ktup* = 2) or 17 min (*ktup* = 1) on a Sun 3/50 workstation. (A) Distribution of initial scores with *ktup* = 2. Three numbers are shown to the left of the histogram: the score reported in the histogram interval (<2, 3–4, . . . , 99–100, > 100); the number of library sequences that obtained an initial similarity score in the histogram interval *(initn)*; and the number of sequences in the library with a best single initial region similarity score in the histogram interval *(init1)*. The *init1* value is identical to the initial score reported by the FASTP program. When there is a difference between the number of library sequences reported in columns two and three, the column two *(initn)* values are graphed with a + and the column three *(init1)* values with a −. The mean of the distribution of *initn* similarity scores was 27.8, with a standard deviation of 7.4. The mean best single *(init1)* similarity score was 27.2 ± 6.1. (B) Distribution of scores with *ktup* = 1. The mean initial score was 36.1 ± 8.9, and the mean *init1* score was 34.6 ± 6.6.

TABLE III

RECEPTOR SEQUENCES RELATED TO β-ADRENERGIC RECEPTOR, *ktup* = 2

	SWISS-PROT entry	Definition	Score		
			initn	*init1*	Optimized
	B2AR$MESAU	β_2-Adrenergic Receptor	2177	2177	2177
	B2AR$HUMAN	β_2-Adrenergic Receptor	1919	1781	1917
	B1AR$MELGA	β_1-Adrenergic Receptor	1140	798	1155
	B1AR$HUMAN	β_1-Adrenergic Receptor	1088	768	794
	ACM3$RAT	Muscarinic acetylcholine receptor M3	429	302	390
	ACM1$RAT	Muscarinic acetylcholine receptor M1	411	260	353
	ACM1$PIG	Muscarinic acetylcholine receptor M1	409	260	353
	ACM2$PIG	Muscarinic acetylcholine receptor M2	330	259	356
	ACM2$HUMAN	Muscarinic acetylcholine receptor M2	330	259	356
10	ACM4$HUMAN	Muscarinic acetylcholine receptor M4	221	183	336
	ACM4$RAT	Muscarinic acetylcholine receptor M4	217	180	334
	OPSR$HUMAN	Red-sensitive opsin (red cone)	128	114	186
	OPSG$HUMAN	Green-sensitive opsin (green cone)	126	117	187
	OPSD$BOVIN	Rhodopsin	108	67	185
	OPS2$DROME	Opsin RH2 (Ocellar opsin)	101	77	256
	OPSD$HUMAN	Rhodopsin	93	71	189
	OPS4$DROME	Opsin RH4 (inner R7 photoreceptor cells)	88	77	211
	SYI$ECOLI	Isoleucyl-tRNA synthetase (isoleucine–RNA ligase)	86	43	47
	CIN2$RAT	Sodium channel protein II, brain	81	61	61
20	OPSB$HUMAN	Blue-sensitive opsin (blue cone)	79	50	152
	NUO5$DROYA	NADH ubiquinone oxidoreductase [NADH dehydrogenase (ubiquinone)]	77	47	58
	A2MG$RAT	α_2-Macroglobulin precursor	76	37	39
	VE1$HPV8	Probable E1 protein	76	52	56
	OPSD$SHEEP	Rhodopsin (fragments)	74	42	62
	CIN3$RAT	Sodium channel protein III, brain	74	61	61
	POLG$POL3L	Genome polyprotein	74	45	45
	Y590$TRYBR	Hypothetical protein C-590	74	51	69
	UDP2$RAT	UDPGlucuronosyltransferase precursor	73	44	44
	CP32$RAT	Cytochrome *P*-450IIIa2 (*P*-450PCN2)	71	51	57
30	UMUA$ECOLI	MucA protein (gene name: *mucA*)	71	58	60
98	OPS1$DROME	Opsin RH1 (outer R1–R6 photoreceptor cells)	63	47	232

TABLE IV
RECEPTOR SEQUENCES RELATED TO β-ADRENERGIC RECEPTOR, *ktup* = 1

	SWISS-PROT entry	Definition	Score		
			initn	*init1*	Optimized
	B2AR$MESAU	β_2-Adrenergic Receptor	2177	2177	2177
	B2AR$HUMAN	β_2-Adrenergic Receptor	1925	1784	1917
	B1AR$MELGA	β_1-Adrenergic Receptor	1155	813	1155
	B1AR$HUMAN	β_1-Adrenergic Receptor	1135	794	794
	ACM3$RAT	Muscarinic acetylcholine receptor M3	483	302	390
	ACM2$PIG	Muscarinic acetylcholine receptor M2	435	250	356
	ACM2$HUMAN	Muscarinic acetylcholine receptor M2	435	250	356
	ACM1$RAT	Muscarinic acetylcholine receptor M1	427	260	353
	ACM1$PIG	Muscarinic acetylcholine receptor M1	425	260	353
10	ACM4$HUMAN	Muscarinic acetylcholine receptor M4	424	247	336
	ACM4$RAT	Muscarinic acetylcholine receptor M4	420	244	334
	OPS1$DROME	Opsin RH1 (outer R1–R6 photoreceptor cells)	190	147	232
	OPS2$DROME	Opsin RH2 (ocellar opsin) (gene name: *RH2*)	186	147	256
	OPS4$DROME	Opsin RH4 (inner R7 photoreceptor cells)	166	83	211
	OPSG$HUMAN	Green-sensitive opsin (green cone)	153	133	187
	OPSD$BOVIN	Rhodopsin	146	102	185
	CIN2$RAT	Sodium channel protein II, brain	136	61	61
	OPSD$SHEEP	Rhodopsin (fragments)	131	64	109
	OPSD$HUMAN	Rhodopsin	131	90	189
20	OPSR$HUMAN	Red-sensitive opsin (red cone)	130	130	186
	RCEM$RHOVI	Reaction center protein M chain	123	69	78
	CIN1$RAT	Sodium channel protein I, brain	122	61	61
	CIN3$RAT	Sodium channel protein III, brain	120	61	61
	NUO4$DROYA	NADH ubiquinone oxidoreductase	112	52	92
	ARG2$CANUT	Arginine metabolism regulation protein II	105	56	56
	POLG$COXB4	Genome polyprotein	103	48	48
	CINA$ELEEL	Sodium chanel protein	102	60	74
	HIP1$YEAST	Histidine permease (gene name: *HIP1*)	100	44	47
	GER2$BACSU	Spore germination protein II (GERA)	99	54	54

TABLE IV *Continued*

	SWISS-PROT entry	Definition	Score		
			initn	*init1*	Optimized
30	CO8B$HUMAN	Complement C8 β chain precursor	98	53	59
	MELB$ECOLI	Melibiose carrier protein	97	47	50
	BGAL$KLEPN	β-Galactosidase	97	51	51
	CYB$DROYA	Cytochrome *b*	95	43	49
	VGV7$ROTAS	Glycoprotein VP7	94	53	53
	COX1$TRYBR	Cytochrome-*c* oxidase polypeptide I	93	55	62
	RCEM$RHOSH	Reaction center protein M chain	93	66	76
	VME1$MHVJH	E1 glycoprotein (matrix glycoprotein)	93	47	47
	Y590$TRYBR	Hypothetical protein C-590	93	65	69
	CP32$RAT	Cytochrome *P*-450IIIa2 (*P*-450PCN2)	93	56	57
40	YCO1$PARDE	Cox locus hypothetical protein 1	92	53	68
45	OPS3$DROME	Opsin RH3 (inner R7 photoreceptor cells)	90	62	142
60	OPSB$HUMAN	Blue-sensitive opsin (blue cone)	86	69	152

Table III also shows the effect of the "joining" step in the calculation of the initial similarity score used for ranking. All the top 20 library sequences have *initn* scores that are greater than the *init1* score, and for several muscarinic acetylcholine receptors the initial FASTA score is higher than the optimized score. For the muscarinic receptor sequences there are two high scoring initial regions that require a gap longer than 32 residues for alignment (Fig. 6), and the lower scoring region is not included in the optimal score.

One can best identify members of the G-protein-coupled receptor family by examining all three of the scores reported by FASTA. The 2 unrelated sequences ranked in the top 20 with *ktup* = 2 [isoleucyl-tRNA synthetase (isoleucine–tRNA ligase) and the sodium channel protein II] have high *initn* scores, but the *init1* scores do not increase when gaps are allowed in the calculation of the optimized score. Likewise, the 98th ranked *Drosophila* opsin is striking because the *init1* score increases almost 5-fold when gaps are introduced. While the *initn* score, which is used to rank the library sequences, provides a very sensitive measure of protein sequence similarity, the relationship between the *init1* score and the optimized score provides a more selective perspective.

The effect of increasing the sensitivity of the search can be seen in Fig.

OPS3$DROME Opsin RH3 (inner R7 photoreceptor cells)
19.8% identity in 243 aa overlap

```
                                          10        20        30        40        50
b2adren                                MGPPGNDSDFLLTTNGSHVPDHDVT-EERDEAWVVGMAILMSVIVLAIVFGNVLVI
                                                  :.:.: .:  :. :.    ...X   ..:  ..:: :::
OPS3$D  MESGNVSSSLFGNVSTALRPEARLSAETRLLGWNVPPEELRHIPEHWLTYPEPPESMNYLLGTLYIFFTLMSMLGNGLVI
              10        20        30        40        50        60        70        80

          60        70        80        90       100       110       120       130
b2adren TAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGASHILMKMWNFGNFWCEFWTSIDVLCVTASIETLCVIAVDRYIAI
         ...  ..:.: .: ....:: :..: .. .:^v . . . ...:.. :.... :.    .:. .:   :: ::. .:
OPS3$D  WVFSAAKSLRTPSNILVINLAFCDFMM-MVKTPIFIYNSFHQGYALGHLGCQIFGIIGSYTGIAAGATNAFIAYDRFNVI
              90       100       110       120       130       140       150

          140       150       160       170       180       190       200       210
b2adren TSPFKYQSLLTKNKARMVILMVWIVSGLTSFLPIQMHWYRATHQKAIDCYHKETCCDFFTNQAYAIASSIVSFYVPLVVM
        :.:.  .. .:..::  .:..... ..       : : ...  ....  .  : :... ..    . ::  : ...
OPS3$D  TRPM--EGKMTHGKAIAMIIFIYMYATPWVVACYTETWGRFVPEGYLTSCTFDYLTDNFDTRLFVACIFFFSFVCPTTMI
      160       170       180       190       200       210       220       230

          220       230       240       250       260       270       280       290
b2adren VFVYSRVFQVAKRQLQKIDKSEGRFHSPNLGQVEQDGRSGHGLRRSSKFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIV
        .. ::..   . .. . . . . ...  .:    ..... ..:
OPS3$D  TYYYSQIVGHVFSHEKALRDQAKKMNVESLRSNVDKNKETAEIRIAKAAITICFLFFCSWTPYGVMSLIGAFGDKTLLTP
        240       250       260       270       280       290       300       310

          300       310       320       330       340       350       360       370
b2adre  HVIQDNLIPKEVYILLNWLGYVNSAFNPLIYCRSPDFRIAFQELLCLRRSSSKAYGNGYSSNSNGKTDYMGEASGCQLGQ

OPS3$D  GATMIPACACKMVACIDPFVYAISHPRYRMELQKRCPWLALNEKAPESSAVASTSTTQEPQQTTAA
        320       330       340       350       360       370       380

          380       390       400       410
b2adren EKESERLCEDPPGTESFVNCQGTVPSLSLDSQGRNCSTNDSPL
```

FIG. 5. Alignment of the β_2-adrenergic receptor and *Drosophila* opsin. The FASTA alignment (*ktup* = 1) of the β_2-adrenergic receptor (labeled b2adren) and a very distantly related opsin (OPS3$DROME, labeled OPS3$D), ranked 45th in Table IV, is shown. The initial score for this comparison is 90, the score of the best single initial region is 62, and the score of the aligned amino acids in the optimized region denoted by ":" and "." is 142. Aligned amino acid identities are denoted by ":"; substitutions with PAM250 scores of zero or greater are denoted by ".".

4B and Table IV, which show the results of a search with *ktup* = 1. In the search with *ktup* = 2, there were only 19 library sequences with *initn* scores greater than 80. With *ktup* = 1, that number increases to 93. Of those 93 sequences, only 21 belong to the G-protein-coupled receptor family (Table IV). However, the *Drosophila* opsin sequence OPS1$DROME, which was ranked 98th with *ktup* = 2, is now ranked 12th, and *Drosophila* opsin sequence OPS3$DROME, which was not found in the top 200 seqeunces with *ktup* = 2, is now ranked 45th. (With *ktup* = 2, the OPS3$DROME similarity scores were: *initn* = 45, *init1* = 34, *opt* = 34.) Nevertheless, the human *mas* oncogene, which is also a member of the G-protein-coupled receptor family, is ranked 203th with *ktup* = 1. Thus, increasing the sensitivity of the search, both by joining initial regions and by looking for

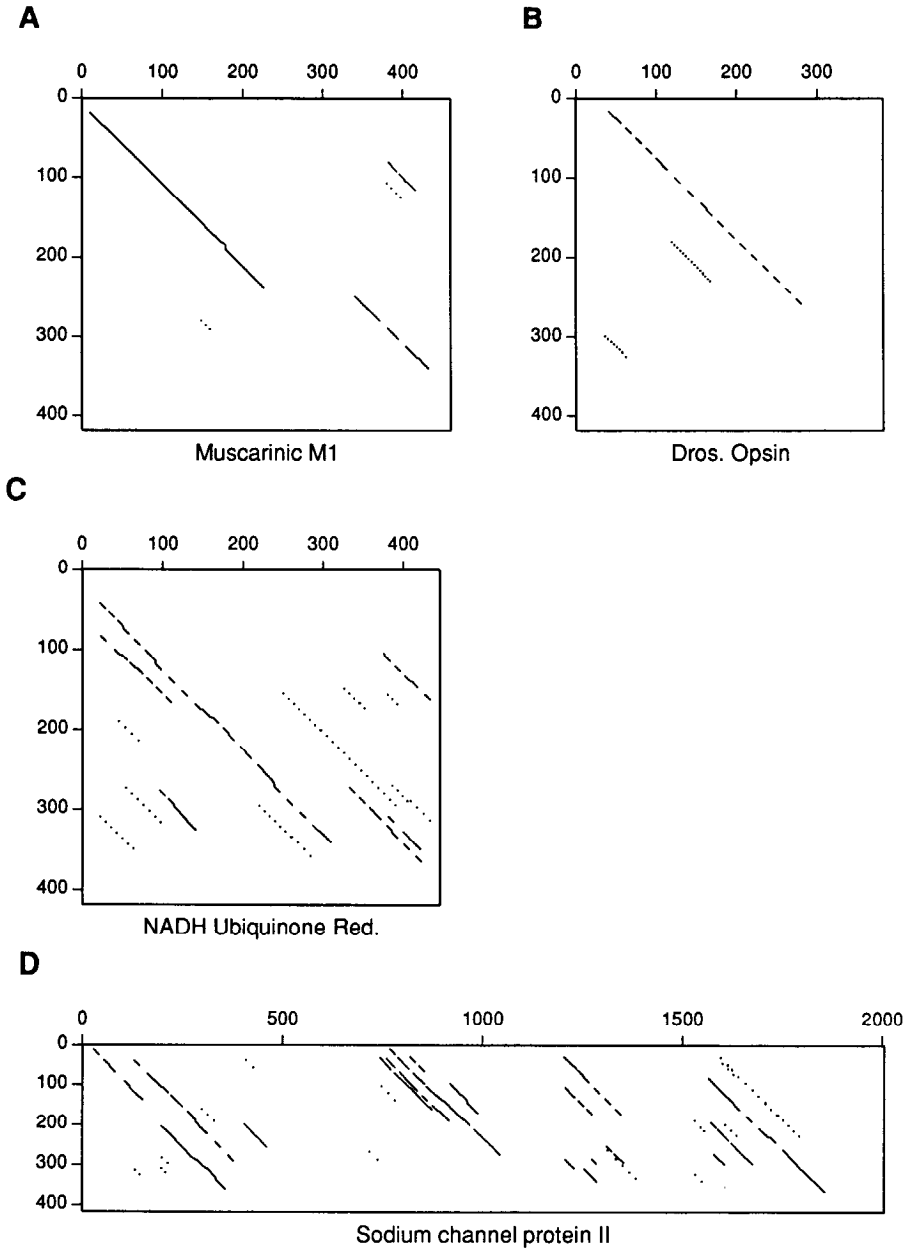

FIG. 6. Local sequence similarities between the β_2-adrenergic receptor and related and unrelated membrane proteins. The β_2-adrenergic receptor was compared with two related G-protein-coupled receptors with the PLFASTA program: (A) the M1 subtype of the muscarinic acetylcholine receptor (SWISS-PROT entry ACM1$PIG) and (B) the *Drosophila* opsin

similarities with *ktup* = 1, raises the scores of distantly related sequences but also dramatically increases the scores of unrelated sequences.

Figure 5 shows the alignment of a distantly related G-protein-coupled receptor, the *Drosophila* opsin found in the inner R7 photoreceptor cells (OPS3$DROME), with the β_2-adrenergic receptor. The optimization procedure uses the standard PAM250 matrix to score identities and substitutions, and it penalizes −12 for the first residue in a gap and −4 for each additional residue in a gap. These gap penalties can be changed by modifying the scoring matrix file (Appendix 2). The amino-terminal boundary of the *init1* initial region is denoted by an "X" at residue 39 in b2adren (Fig. 5). The carboxy-terminal boundary of this region is denoted by a "^" at residue 89 in b2adren and a "v" at residue 114 in OPS3$D. These two residues were aligned without gaps when the region was identified in the initial FASTA comparison, but the introduction of the gap after residue 117 in OPS3$D caused the two phenylalanines to be offset from one another. The best initial region found with *ktup* = 1 includes transmembrane regions I and II of the β_2-adrenergic receptor and *Drosophila* opsin; optimization extends the alignment to include transmembrane regions III, IV, and V. The alignment of regions II and III reveals that the *Drosophila* opsin has several residues that are conserved in almost all members of the G-protein-coupled receptor superfamily in the appropriate relative positions (Leu^{75}, Asp^{79}, Arg^{131}).

In this example, FASTA is able to align only five of the seven membrane-spanning regions in the two receptors. However, FASTA can align all seven of the transmembrane regions of the β_2-adrenergic receptor with another *Drosophila* opsin, OPS4$DROME, which shares 72% identity with OPS3$DROME. FASTA will align all seven transmembrane regions of the β_2-adrenergic receptor with OPS3$DROME if the penalty for each residue in a gap is decreased from −4 to −2.

(Swiss-Prot entry OPS3$DROME) shown in Fig. 5. Also shown are similarities to (C) NADH ubiquinone oxidoreductase and (D) a high-ranking but unrelated rat sodium channel protein II (SWISS-PROT entry CIN2$RAT). The vertical scale refers to the β_2-adrenergic receptor sequence in each case. Lines are drawn to indicate the residues aligned in the local similar regions found by the LFASTA program. The sequences were compared with *ktup* = 1. (A) The aligned region denoted by the solid line from residue pair 11, 20 (ACM1$PIG, b2adren) to 227, 238 has an initial score of 260 and an optimized score of 353. The region denoted by the offset diagonal broken line in the lower right corner (341, 250 to 340, 433) has an initial score of 139 and an optimized score of 173. The other off-diagonal regions have initial scores ranging from 38 to 56 and optimized scores ranging from 39 to 63. (B) The region denoted by the dashed line near the central diagonal (11, 20 to 227, 238) has an initial score of 62, and an optimized score of 140. (D) Eleven similar regions are shown, with initial scores ranging from 35 to 61, and optimized scores ranging from 35 to 90. The region from 130, 39 to 378, 290 has the highest optimized score, 90, and an initial score of 39.

The partial alignment (Fig. 5) includes two additional values that can be used to evaluate the significance of the sequence similarity: the percent identity and the length of the aligned region. In general, the length of the aligned region is a better indicator of significance than the actual percent identity. It is not uncommon to find short regions of sequences (15–40 amino acids) that share 30–50% identity in unrelated proteins. However, sequences that share more than 20–25% identity over their entire length almost always share a common ancestor, and it is possible to show convincingly that sequences which share as little as 15% identity over their entire length are homologous.

Evaluating Sequence Similarities: RDF2 and LFASTA

Examination of the 40 top scoring sequences in Table IV suggests that the β-adrenergic receptor shares strong similarity, and almost certainly common evolutionary ancestry, with the muscarinic acetylcholine receptors, rhodopsin, and the opsins.[17] In addition, many of the highly ranked sequences that do not belong to the G-protein-coupled receptor family are membrane proteins, which presumably share sequence similarity because of the requirement to form structures that span membranes. Often, these high ranking but unrelated library sequences have high *initn* scores but much lower *init1* scores, and the *init1* scores increase very little, if at all, when the band around the best initial region is optimized. However, if the G-protein-coupled receptors were not in the library, one might propose that sodium channel protein II or NADH ubiquinone oxidoreductase [NADH dehydrogenase (ubiquinone)] share common ancestry with the β_2-adrenergic receptor. The sodium channel protein is tempting because it has such a high initial score, but one must be cautious because the *init1* score does not increase with optimization. The NADH ubiquinone reductase is even more tantalizing because it has a high *initn* score, and its *init1* score almost doubles with optimization.

The RDF and RDF2 programs test whether high similarity scores are likely to reflect sequence similarity that is due to common ancestry or simply a locally biased amino acid composition. These programs compare two sequences, calculating initial and optimized scores, and then shuffle the second sequence a specified number of times (100 to 200 shuffles are recommended), again calculating the initial and optimized scores. The earlier RDF program shuffled the second sequence by moving each of its residues to a random position in the shuffled sequence, thus preserving the length and amino acid composition of the sequence, and calculated *init1*

[17] R. A. F. Dixon, B. K. Kobilka, D. J. Strader, J. L. Benovic, H. G. Dohlman, T. Frielle, M. A. Bolanowski, C. D. Bennett, E. Rands, R. E. Diehl, R. A. Mumford, E. E. Stater, I. S. Sigal, M. G. Caron, R. J. Lefkowitz, and C. D. Strader, *Nature (London)* **321**, 75 (1986).

and optimized scores using the FASTP procedure. With this shuffling procedure, however, local regions of sequence bias are spread throughout the shuffled sequence. The RDF2 program provides a more stringent shuffling procedure in which the second sequence is shuffled in short blocks (usually 10 residues). With RDF2, residues 1–10 of the original sequence become residues 1–10 of the second sequence, but in a different order, and so on for 21–30, 31–40, etc. A local shuffling procedure is particularly appropriate for examining sequences where local sequence bias is expected, as in the membrane-spanning helices of membrane proteins. RDF2 then calculates three similarity scores for each shuffled sequence using the FASTA procedure.

RDF2 was used to evaluate the similarities between the β_2-adrenergic receptor and either members of the G-protein-coupled receptor family or other unrelated but high scoring membrane proteins (Table V). RDF2 provides several perspectives from which one can evaluate the statistical significance of a similarity score. One perspective is the z value, which is calculated by subtracting the mean score of the randomly shuffled sequences from the score of the unshuffled sequence and then dividing by the standard deviation of the distribution of shuffled scores. RDF2 calculates z values for each of the three similarity scores calculated by FASTA, but the z value for the optimized score is the most informative (Table V). For the example of the β_2-adrenergic receptor versus *Drosophila* opsin OPS3$DROME, the optimized score is 142, the mean optimized score for the shuffled sequences is 50.3, and the standard deviation of the distribution of optimized shuffled sequence scores is 10.4. Therefore, the unshuffled score is 8.8 standard deviations above the mean, or has a z value of 8.8. In an earlier paper,[1] we suggested that one should be skeptical of conclusions based on sequence similarity scores with z values less than 3, and more confident when the z values are greater than 6. The z values in Table V tend to support this guideline: the lowest scoring member of the G-protein-coupled receptor family has a z value of 4.3, and the highest scoring unrelated sequence has a z value of 2.9.

Unfortunately, the z values determined in a Monte Carlo analysis become less informative as the distribution of similarity scores diverges from a normal distribution. Since FASTA searches have a more asymmetric distribution of scores than FASTP ones, an alternative perspective focuses on the highest scores of the shuffled sequences (the maximum column in Table V). For example, after 200 shuffles, the highest optimized similarity score between the β_2-adrenergic receptor and a shuffled copy of the NADH ubiquinone reductase was 100; 3 of the 200 shuffled sequences obtained optimized similarity scores greater than the value of 92 found for the unshuffled sequence. In contrast, the highest optimized similarity score

TABLE V

STATISTICAL SIGNIFICANCE OF SIMILARITY SCORES[a]

Shuffled Sequence	*initn* score				Optimized score			
	Unshuffled	Average	Maximum	z value	Unshuffled	Average	Maximum	z value
Drosophila opsin OPS4$DROME	166	51.5 ± 14.9	102	7.7	211	48.8 ± 9.9	96	16.3
Drosophila opsin OPS3$DROME	90	52.3 ± 13.8	94	2.7	142	50.3 ± 10.4	93	8.8
Human opsin (blue-sensitive)	86	56.3 ± 16.0	113	1.8	153	50.7 ± 9.5	90	10.7
Human *mas* oncogene	75	55.0 ± 15.0	101	1.3	101	53.1 ± 10.9	122	4.3
NADH ubiquinone oxidoreductase NUO4$DROYA	112	67.5 ± 18.8	127	2.4	92	59.6 ± 11.0	100	2.9
Rat sodium channel CIN2$RAT	136	70.5 ± 18.0	118	3.66	61	54.0 ± 9.8	93	0.7
Eel sodium channel CINA$ELEEL	102	78.3 ± 18.5	122	1.29	74	59.7 ± 12.3	112	1.2

[a] RDF2 was used to compare the hamster β_2-adrenergic receptor with the sequences listed using *ktup* = 1. Each sequence was shuffled 200 times, using a local shuffle with a window of 10 residues.

for the shuffled *Drosophila* opsin sequence was 93, substantially lower than the optimized score of 142 for the unshuffled sequence. Thus, the *Drosophila* opsin sequence appears to be related to the β_2-adrenergic receptor, and the NADH ubiquinone reductase unrelated, by both the criteria of z values and the highest shuffled scores.

Another way to evaluate sequence similarity is to compare the sequences with LFASTA and PLFASTA. While FASTA reports a similarity score based on the alignment of multiple initial regions that do not overlap one another and displays the one best alignment between the two sequences, LFASTA displays all the local regions of similarity shared between two sequences. LFASTA displays the actual local alignments and similarity scores, and PLFASTA plots the alignments in a form similar to a dot matrix plot (Figs. 6 and 7). Plots of local similarities between the β_2-adrenergic receptor and four related and unrelated sequences are shown in Fig. 6. The PLFASTA comparison of the β_2-adrenergic receptor with the muscarinic receptor shows why the initial similarity score for this comparison is higher than the optimized score; there are two regions with strong similarity that are shared between these two sequences, but they are separated by an insert of 70 amino acids in the muscarinic acetylcholine receptor sequence. These two regions of similarity are joined in the calculation of the *initn* initial similarity score by FASTA. Both of the protein sequences that are related to the β_2-adrenergic receptor show a strong region of similarity along a single major diagonal, with a few weak off-diagonal regions. In contrast, the NADH reductase and the sodium channel have a large number of regions with modest similarity scores. In the case of the sodium channel protein, four distinct regions of similarity are found. These regions correspond to the four clusters of six transmembrane segments that have been predicted.[18] Although none of the local regions of similarity in the unrelated sequences has a very high score, there are so many of them that they can be joined together. This large number of similar regions accounts for the high initial similarity scores calculated for these sequences as well as the much lower *init1* and optimized similarity scores.

LFASTA, PLFASTA, and Local Similarity

LFASTA and PLFASTA can also be used to examine repeated structures in proteins, or to map exons in genomic clones using cDNA sequences (see Fig. 2 in Ref. 3). LFASTA used the same first two steps for finding initial regions that FASTA uses. Instead of saving ten initial re-

[18] W. A. Catterall, *Science* **242**, 50 (1988).

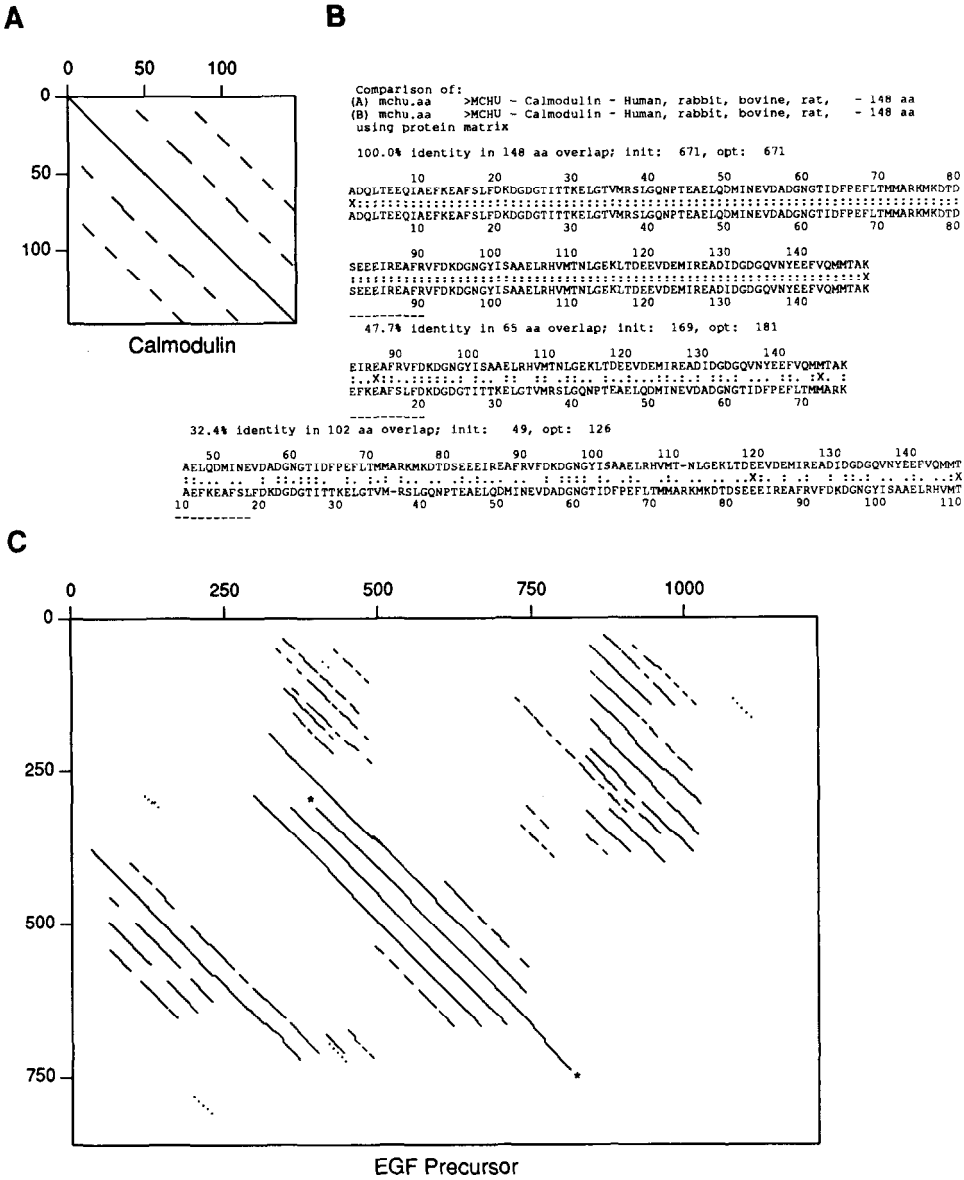

FIG. 7. Examination of local sequence similarities with LFASTA and PLFASTA. (A) The PLFASTA program was used to compare the human calmodulin sequence (PIR entry MCHU) with itself (*ktup* = 2). The ordinate and abscissa indicate the position in the calmodulin sequence. (B) Sample output from the LFASTA program comparing calmodulin to itself. Only two of the four off-diagonal alignments are shown, since the other two are symmetrically identical. As in Fig. 5, "X" denotes the boundaries of the initial region found by

gions, however, LFASTA saves all diagonal regions with similarity scores greater than a threshold. In addition, instead of focusing on a single region, LFASTA computes a local alignment for each initial region. The ends of the optimized local region of similarity are determined by scanning backward, then forward, in the sequences. Starting at the end of the initial region, a Smith and Waterman type of optimization[15] is performed in a band that is centered around the initial region. The optimization continues past the beginning of the initial regions until all possible alignment scores have gone to zero. The location where the maximal local similarity score occurred during the backward scan is saved and used as the starting position of a second optimization that proceeds in the forward direction. The forward optimization proceeds in the same way until all possible alignment scores have gone to zero; then the position where the new maximal local similarity score occurred is saved. An optimal path starting at that maximum is then generated by a standard traceback procedure.[10] In some cases, because of the dynamic boundaries of the optimization, several initial regions may be included in a single optimization. A check is therefore made to prevent the output of identical alignments.

Figure 7 shows two examples of how LFASTA and PLFASTA can be used to examine internal duplications and repeated domains in protein sequences. Figure 7A,B displays duplications within the human calmodulin sequence. Calmodulin is a member of a superfamily of calcium-binding proteins that share an E-F hand structure; different members of this family contain from two to six E-F hand calcium-binding domains. Calmodulin contains four E-F hand domains, which can be numbered 1 through 4. The similarity scores and percent identities determined by LFASTA (Fig. 7B) suggest that domains 1 and 2 diverged from 3 and 4 more recently (they share 48% identity) than 1 diverged from 2 or 3 from 4.

Figure 7C shows the complex pattern of exon-shuffling that apparently took place during the evolution of the LDL receptor and the EGF precursor.[4] The four solid diagonal lines that align residues 250–750 of the LDL receptor with residues 300–750 of the EGF precursor represent alternative alignments of an EGF precursor domain that is present three times in the

LFASTA. (C) The PLFASTA program was used to compare the human LDL receptor (PIR entry QRHULD) and the mouse EGF precursor (PIR entry EGMSMG) with *ktup* = 1. The position in the LDL receptor is shown on the ordinate; the position in the EGF precursor is shown on the abscissa. The highest scoring local alignment, indicated with an asterisk, has an initial score of 274 and an optimized score of 654. The other solid diagonal lines indicate regions with optimized similarity scores ranging from 252 to 350, the long-dashed diagonal lines show regions with scores ranging from 110 to 144, the short-dashed lines show regions with scores between 50 and 100, and the dotted lines indicate regions with scores less than 50.

LDL receptor and four times in the EGF receptor. The diagonal indicated by the asterisk would be the only alignment reported by FASTA.

TFASTA and DNA Library Searches

Included in the FASTA package of programs is TFASTA, a program for comparing a protein sequence to a DNA sequence by translating the DNA sequence in all six reading frames. The value of doing sequence searches with protein rather than DNA sequences cannot be overemphasized. A search of the mammalian portion of GenBank (Release 58) with the mRNA sequence encoding the hamster β_2-adrenergic receptor (GenBank locus HAMARBR, 2015 nucleotides), which required 35 min on a Sun 3/260 computer, did not detect any G-protein-coupled receptors except other β-adrenergic and muscarinic acetylcholine receptors. In contrast, a search of the same database with TFASTA and *ktup* = 2 took 17 min and revealed the mammalian G-protein-coupled receptors shown in Table III. Since TFASTA compares a protein to a DNA sequence in all six reading frames, it can also be used to check for frameshifts in cDNA sequences when other homologs for the protein coded by the cDNA are known.

Short Sequences

Although FASTA was originally designed to search protein sequence libraries for homologous sequences, it can also be used to search for oligonucleotides and oligopeptides. Nevertheless, a few adjustments to the normal search parameters may be required for satisfactory results. When searching with short query sequences, the *ktup* parameter should almost always be set to 1. In addition, care must be taken to make certain that the CUTOFF score is not set too high by default. The CUTOFF score is a value that is used to decide whether a particular library sequence should be saved for later display and optimization. The CUTOFF score is calculated by a predetermined formula based on the length of the sequence and the *ktup* value. If the sequence is so short that the CUTOFF score is less than three times the length of the sequence, the program warns that the CUTOFF value may be too high and prompts the user for a new value. Unfortunately, this value may be inappropriate for DNA sequences. In these cases, the program may not prompt even though the CUTOFF value is too high, with the result that even though all the library is scanned, no sequences are saved. To prevent this from happening, the CUTOFF value can be set to a very low number such as 5 by using the symbol CUTOFF. In addition, the default values in the DNA scoring matrix may be too conservative for oligonucleotide sequences (+4 for an identical match, −3 for a mismatch,

+2 for a match to an ambiguous base, −12 for the first residue in a gap, −4 for additional residues).

Searching with Different SMATRIX

It is possible to use other scoring matrices with the FASTA programs, and files for alternative matrices are included with the program package. To use an alternative matrix, one can either define the environment symbol SMATRIX to the name of the file that contains the alternative matrix or enter the alternative SMATRIX file name on the command line preceded by "−s." By setting SMATRIX to a different matrix file at the beginning of a work session, one can use the same scoring parameters for FASTA, RDF2, or LFASTA. However, setting SMATRIX can cause FASTA to treat a DNA sequence as a protein sequence, or vice versa, so one can also change the SMATRIX by including its file name on the command line using the "−s" option. For example,

```
FASTA −s codaa.mat
```

will cause FASTA to use a matrix based on the genetic code. Our experience has been that different scoring matrices can change the sensitivity of a FASTA search (most matrices are less sensitive than the PAM250 matrix), but that the relative ranking of the library scores remains about the same. Alternative scoring matrices and modification of the gap penalties have more effect on the precise sequence alignments than on relative similarity scores.

Considerable flexibility has been built into FASTA, so that virtually every aspect of the search process can be modified. For example, line 6 in the scoring matrix file (Appendix 2) specifies the relationship between the residues of a sequence and the number used in the lookup table calculation. Thus, during the lookup calculation, it is possible to cause several different residues to have the same value and appear to be the same. This mapping of residues to lookup values is used for DNA searches so that ambiguous residues such as R (purine, A or G) or Y (pyrimidine, C or T) can match sequences with no ambiguity codes. In this case, an R in the query or library sequence would match an A or R (but not a G), and a Y would match a C or Y. While this mapping method is not perfect, since a match between an R and G is not found during the first step of the search (the R-G match would be scored correctly in the second rescanning step), it allows ambiguous codes to be recognized part of the time even when they are rare in the library.

This mapping for the lookup table can also be used to change the way FASTA looks for initial regions within protein sequences. For example, one might classify amino acids into six groups: acidic, basic, small side

chain, large hydrophobic side chain, aromatic, and cysteine. Each of the amino acids could then be placed in a group and FASTA would search for initial regions with high densities of "identities," but any member of the same group would be considered an "identity."

A similar strategy can be used to scan a set of consensus DNA binding sites. FASTA used the IUPAC–IUB code for ambiguous nucleotides (Appendix 4), which allows all 15 possible nucleotides and ambiguities to be specified. Thus, one could make a library of known transcription factor binding sites and compare binding sites for a newly characterized factor by comparing the new binding site to the library. Some judgment would have to be used in encoding of the ambiguous residues, and FASTA cannot distinguish between an A–R match at one position and an A–R match at a different position (although a modified version of FASTA has been developed that does allow position-specific scoring). Searches with short sequences against libraries with large numbers of ambiguous residues should always be done with *ktup* = 1.

Output Options

FASTA also includes a variety of other input and output options, which are listed in Appendix 3. Several output options control how much of the sequence alignment is shown, how identities and substitutions are highlighted, and the number of residues displayed on each line. However, one must remember that FASTA highlights only identities and substitutions in the aligned region of the two sequences, and this region may not contain some residues at the ends of the sequences. In addition, the "–Q" (quiet) option can be used to allow the FASTA program to run without requesting any additional input. This option allows one to do FASTA searches in the background on some computers, or to do several FASTA searches, one after another, on an IBM-PC.

Summary

The FASTA program can search the NBRF protein sequence library (2.5 million residues) in less than 20 min on an IBM-PC microcomputer and unambiguously detect proteins that shared a common ancestor billions of years in the past. FASTA is both fast and selective because it initially considers only amino acid identities. Its sensitivity is increased not only by using the PAM250 matrix to score and rescore regions with large numbers of identities but also by joining initial regions. The results of searches with FASTA compare favorably with results using NWS-based programs that are 100 times slower. FASTA is slightly less sensitive but considerably more selective. It is not clear that NWS-based programs would be more successful in finding distantly related members of the

G-protein-coupled receptor family. The joining step by FASTA to calculate the *initn* score is especially useful for sequences that share regions of sequence similarity that are separated by variable-length loops.

FASTP and FASTA were designed to identify protein sequences that have descended from a common ancestor, and they have proved very useful for this task. In many cases, a FASTA sequence search will result in a list of high scoring library sequences that are homologous to the query sequence, or the search will result in a list of sequences with similarity scores that cannot be distinguished from the bulk of the library. In either case, the question of whether there are sequences in the library that are clearly related to the query sequence has been answered unambiguously. Unfortunately, the results often will not be so clear-cut, and careful analysis of similarity scores, statistical significance, the actual aligned residues, and the biological context are required. In the course of analyzing the G-protein-coupled receptor family, several proteins were found that, because of a high *initn* score and a low *init1* score that increased almost 2-fold with optimization, appeared to be members of this family which were not previously recognized. RDF2 analysis showed borderline z values, and only a careful examination of the sequence alignments that focused on the conserved residues provided convincing evidence that the high scores were fortuitous. As sequence comparison methods become more powerful by becoming more sensitive, they become more likely to mislead, and even greater care is required.

Appendix 1. FASTA File Formats

FASTA and TFASTA can search library files in the following formats on non-VAX/VMS systems:

(1) FASTA/DM (query sequence), library type 0

```
>SEQID1 - title line
either protein sequence or DNA sequence
>SEQID2 - comment line
AGTHKPRY...
```

(2) GENBANK tape format, library type 1

```
LOCUS        HUMHBB  ....
DEFINITION   ....
ORIGIN       .....
        1  ACGT....
```

The GENBANK DNA sequence library is available on tapes in this format from:

GenBank
c/o IntelliGenetics, Inc.
700 El Camino Real East
Mountain View, CA 94040

(3) PIR Codata format (library type 2)

```
ENTRY           CCHU        #Type Protein
TITLE           Cytochrome c - Human
SEQUENCE
                5 ...
      1 A F T G H I E W ...
```

The NBRF/PIR protein sequence library is available in this format from:

Protein Identification Resource
National Biomedical Research Foundation
Georgetown University
3900 Reservoir Rd., N. W.
Washington, D.C. 20007

(4) EMBL/SWISS-PROT Format, library type 3

```
ID   16K$TRVPS      STANDARD;      PRT;   141 AA.
DE   16 KD PROTEIN.
SQ   SEQUENCE   141 AA;  16297 MW;  93420 CN;
     DVYNCCGRSH LEKCRKRVEA RNREIWKQIR RIQAESSSAT RKKSHNSKNS KKKFKEDREF
```

DNA and protein sequence libraries in this format are distributed by:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10 2209
D-6900 Heidelberg
Federal Republic of Germany

(5) IntelliGenetics format, library type 4

```
;comment
;comment
SEQID
ABCDEFG...
```

(6) GenBank compressed floppy disk format, library type 9. Files in this format containing the GenBank DNA sequence library are distributed on IBM-PC and Macintosh floppy disks by IntelliGenetics.

On VAX/VMS systems, the FASTA programs can read sequences in the NBRF/PIR VAX/VMS file format, and the University of Wisconsin Genetics Computer Group format.

Appendix 2. SMATRIX file

An sample SMATRIX file for DNA sequences is shown. The line numbers are referred to in the text below.

```
;D standard DNA scoring matrix                                    1
 1 45 80 5 6 80 4                                                 2
 -12 -4                                                           3
 * @ 0 1 2                                                        4
 A  C  G  T  R  Y  M  W  S  K  D  H  V  B  N                      5
 0  1  2  3  0  1  0  0  1  2  0  0  0  1  0                      6
 4                                                                7
-3  4
-3 -3  4
-3 -3 -3  4
 2 -1  2 -1  2
-1  2 -1  2 -2  2
 2  2 -1 -1  0  0  2
 2 -1 -1  2  0  0  0  2
-1  2  2 -1  0  0  0  0  2
-2 -2  1  1  0  0  0  0  0  2
 1 -2  1  1  1  0  0  1  0  1  1
 1  1 -2  1  0  1  1  1  0  0  0  1
 1  1  1 -2  1  0  1  0  1  0  0  0  1
-2  1  1  1  0  1  0  0  1  1  0  0  0  1
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

1. `;P` or `;D`, this comment, if present, is used to determine whether sequences should be labeled as amino acids (aa) or nucleotides (nt).

2. Scoring parameters:

   ```
   FACT BESTOFF BESTSCALE BKFACT BKTUP BESTMAX HISTSIZ
   ```

   KFACT is used in the "diagonal method" search for the best initial regions, for proteins, KFACT = 4, for DNA, KFACT = 1 (used only if PAMFACT=0).

   BESTOFF, BESTSCALE, BKFACT, BKTUP and BESTMAX are used to calculate the cutoff score. The bestcut parameter is calculated from parameters 2 - 6. If N0 is the length of the query sequence:

   ```
   BESTCUT = BESTOFF + N0/BESTSCALE + BKFACT*(BKTUP-KTUP)
   if (BESTCUT>BESTMAX) BESTCUT=BESTMAX
   ```

   HISTSIZ is the size of the histogram interval.

3. Deletion penalties. The first value is the penalty for the first residue in a gap, the second value is the penalty charged to each subsequent residue in a gap.

4. End of sequence characters. These are not required, since FASTA knows how to find the beginning of a library sequence, but they can be used if sequences have additional comments after the end. If not used, the line must be left blank.

5. The alphabet. The program automatically converts upper to lower case and vice-versa.

6. The lookup table values for each letter in the alphabet. This allows several characters to be hashed to the same value, e.g. a DNA sequence alphabet with A = adenosine, R = purine, N = any base, would have each of these characters treated as 0. The lowest hash value should be 0.

7ff. The lower triangle of the symmetric scoring matrix. There should be exactly as many lines as there are characters in the alphabet, and the last line should have n-1 entries. The program does not check for the length of each line, so it is easy to use an incorrect matrix by having fewer entries in the scoring matrix than in the alphabet.

Appendix 3. FASTA Options

Scoring parameters, output line lengths, and other features of FASTA can be modified either by setting an environment symbol or on the command line. For example, to have alignments be displayed with 80 residues per line, one can either set an environment variable:

```
set LINELEN=80 (PC-DOS)   or  FASTA -l 80
```

Command Line Option	Environment Symbol	Function
-a	SHOWALL=1	Normally the optimized region of an alignment is shown in context, but the complete sequence may not be shown if the optimized region does not extend near the end of the sequence. With this option, complete sequences are always shown.
-c #	CUTOFF=#	The CUTOFF value (#) is the threshold for saving a sequence in a list of sequences to be sorted and optimally aligned after the search. This value is also used as the threshold for the optimal alignment of initial regions in the second step of FASTA.
-f	PAMFACT=1	Use the newer FASTA variable SMATRIX score for a *ktup* match. Default for protein comparisons.
-k	PAMFACT=0	Use a constant (FASTP) score in scan for a *ktup* match. Default for DNA.
-l file	FASTLIBS	File name for the location of library menu file.

-m #	MARKX = #	(0, 1, 2) MARKX modifies the way aligned residues are highlighted.
		`MARKX=0 (default)  MARKX=1  MARKX=2`
		`MWRTCGPPYT  MWRTCGPPYT  MWRTCGPPYT`
		`::..:: :::  xx X  ..KS..Y...`
		`MWKSCGYPYT  MWKSCGYPYT`

-o #		The number of scores and alignments to be reported by default. (Used in conjunction with -Q).
-p #	GAPPEN	The gap threshold for joining two initial regions in the calculation of the *initn* score. Normally set to the CUTOFF value.

-Q		Quiet mode - FASTA does not prompt for any input. The default number of scores and alignments are displayed on the terminal or written to the standard output file. This option is used for running in batch mode, or in the background.
-s *file*	SMATRIX	The scoring matrix is read from *file*.
-w #	LINLEN=#	Number of residues per line for sequence alignments. This value must be less than 200.
-3		(TFASTA only) translate only three forward frames.

Not all of these options are appropriate for all of the programs. The options above are used by FASTA and TFASTA. RDF2 uses -c, -f, -k, and -s.

Appendix 4. Codes for Ambiguous Nucleotides*

Code	Nucleotide
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil, treated as T
R	Purine, A or G
Y	Pyrimidine, C or T
M	A or C
W	A or T
S	C or G
K	G or T
D	A, G, or T (not C)
H	A, C, or T (not G)
V	A, C, or G (not T)
B	C, G, or T (not A)
N	A, G, C, or T

* FASTA also recognizes X.