



**Dr Clare Sansom**

works part time at Birkbeck College, London, and part time as a freelance computer consultant and science writer. At Birkbeck she coordinates an innovative graduate-level Advanced Certificate course, 'Principles of Protein Structure', which is taught entirely using the Internet.

**Keywords:** database searching, DNA sequences, protein sequences, scoring matrices, BLAST, FASTA

# Database searching with DNA and protein sequences: An introduction

Clare Sansom

Date received (in revised form): 12th November 1999

## Abstract

This review of sequence database searching aims to set out current practice in the area, in order to give practical guidelines to the experimental biologist. It describes the basic principles behind the programs and enumerates the range of databases available in the public domain. Of these, the most important are the equivalent DNA databases European Molecular Biology Laboratory (EMBL), GenBank and DNA Databank of Japan (DDBJ), and the protein databases Swiss-Prot and TrEMBL. The commonly used BLAST and FASTA algorithms are described in detail and alternative approaches mentioned briefly. Scoring matrices used to compare amino acid types during protein database searches are compared, with an emphasis on the PAM and BLOSUM series of observed substitution matrices.

## Introduction

All biological scientists now have access to vast archives of genetic data via the world-wide Internet.<sup>1</sup> Technological advances and public and private investment have led to an explosion in genome sequencing, and a large proportion of sequence data is publicly available. Therefore, biologists seeking to answer the question 'What sequences (in the databases) are most similar to, or contain the most similar regions to, my previously uncharacterised sequence?' are increasingly able to find a statistically significant answer. 'Search engines' have been developed to help answer this question. The basic principle of each of these algorithms is the same. The test sequence is compared with each sequence in a database in turn to establish the 'best scoring' alignment, and those alignments with the highest scores are reported.

Programs for database searching differ in the core algorithm they employ. This affects their speed and sensitivity. Some algorithms, particularly those that have been optimised for speed, use simplified assumptions in scoring sequence similarity, and so may miss

marginal, but still significant, matches.

The time taken by a search depends on the length of the sequence and the size of the database: since databases are increasing in size as computers increase in speed, the speed of the algorithms is still an important consideration. The most widely used programs are implemented on many web servers world-wide. These differ in terms of the choice of databases and those parameters that the user is allowed to change. Most programs can be applied to both DNA and protein sequences (Table 1).

It is preferable to search sequence databases at the protein sequence level if appropriate, as there is a higher 'signal-to-noise' ratio. Quite simply, the 'alphabet' of amino acid types is 20 characters long, whereas the alphabet of bases is only four characters long. Also, the fact that some changes of amino acids are more likely to occur than others can yield useful information. Even if the exact protein translation of a DNA sequence is not known, if the sequence is assumed to represent a coding region it is possible to translate it automatically in all six reading frames

## Sequence databases, Internet

Dr Clare Sansom  
Department of Crystallography,  
Birkbeck College,  
London, WC1E 7HX UK

Tel: +44(0) 20 7631 6800  
Fax: +44(0) 20 7631 6803  
E-mail:  
c.sansom@mail.cryst.bbk.ac.uk

**Table 1:** Comparison of some of the most popular programs for searching sequence databases

Program	Sensitivity	Speed	Sequence types
BLAST	Fairly sensitive	Very fast	DNA, protein
FASTA	Sensitive	Fast	DNA, protein
Blitz	Very sensitive	Fairly fast	DNA, protein
SSEARCH	Very sensitive	Slower	DNA, protein
PSI-BLAST	Extremely sensitive	Slower	Protein

**DNA sequence databases, EMBL, GenBank, DDBJ**

and search a protein database with each translation in turn. This calculation is incorporated into variants of the BLAST and FASTA programs described later.<sup>2-6</sup>

**Choice of databases**

Choosing an algorithm is only one aspect of the design of a database mining strategy. The importance of the choice of an appropriate database is self-evident.

**DNA sequence databases**

There are three equivalent primary databases of generic DNA sequence data: European Molecular Biology Laboratory (EMBL), GenBank<sup>7</sup> and DNA Databank of Japan (DDBJ). The only significant difference between these is their location. The EMBL database is compiled and maintained at the European Bioinformatics Institute in Hinxton; GenBank is based at the National Center for Biotechnology Information in Bethesda, USA, and DDBJ is located in Mishima, Japan. Sequences sent to one database are indexed and distributed automatically to the others. These databases are huge: release 61 of the EMBL database, dated 3rd December 1999, contains 5,303,436 sequence entries comprising 4,508,169,737 nucleotides. This represents an increase of about 27 per cent over release 60, dated September 1999. Currently, they are doubling in size approximately every nine or ten months.

It will not always be necessary, or even desirable, to search the whole of one of the main databases. For example,

users may restrict a search to a particular organism type (such as vertebrates, rodents or prokaryotes). Expressed sequence tags (ESTs) are small fragments of complementary DNA. It may be useful to restrict a database search by excluding ESTs – over 60 per cent of entries in release 60 of EMBL are ESTs – or, alternatively, to search a database consisting entirely of ESTs. Many complete genomes are now available on the web, and it is possible to search a database containing the full gene sequence of a single organism.

**Protein sequence databases**

Many protein sequence databases are very well annotated and information-rich; the entries in these databases are cross-linked to many other databases and information sources. The most frequently used of these databases are undoubtedly Swiss-Prot<sup>8</sup> and TrEMBL. Swiss-Prot contains a relatively small number (83292 at January 2000) of well-annotated protein sequences. TrEMBL, in Swiss-Prot format, is prepared automatically from the coding regions of the EMBL DNA sequence database. A similar database (GenPept) is produced by translating all coding regions of the DNA sequences in GenBank. Some specialist protein databases are also widely used. Two examples are the NRL-3D database, which contains sequences only of those proteins with three-dimensional structures in the Protein Data Bank,<sup>9</sup> and the Kabat database of protein sequences of immunological importance.<sup>10</sup>

**Protein sequence databases, Swiss-Prot, TrEMBL**



## METHODOLOGY

### Statistical significance

It is important to realise that any database search will extract close matches based on calculated similarity between strings of letters. Biologists are likely to be interested in extracting those sequences that can be assumed from similarity to be evolutionarily related to their test sequence. Such sequences, which will have derived from a common ancestor, are defined to be 'homologous'. Extracting this information from a purely numerical measure of similarity is difficult. The most practicable simple guide to the likelihood of a 'hit' in a database search being evolutionarily related to a test sequence is a statistical measure of how likely the match is to have occurred purely by chance. Such values are calculated and quoted in the results generated by database searching algorithms.

The most common measure of probability is the so-called Expect value, or *E*-value. This is the number of alignments with a given score that would be expected to occur at random in the database that was searched. Thus, an *E*-value of 1.00 for a match between a database sequence and a test sequence would indicate that exactly one random sequence in a database of that size would be likely to match the test sequence as well as the current one. *E*-values are independent of the lengths of the sequences. Values as low as  $10^{-50}$  are not uncommon in well-conserved families. With large databases, values between about 0.01 and 10 can be said to represent a 'grey area'; it may be useful to analyse sequences matching at this level in more detail.

Almost all sequence alignment programs – of which programs for database searching are a subset – use a 'scoring' approach. Each position of each alignment is given a score, which is positive for a good match and negative for either a poor match or a

position where a residue in one sequence is matched by a gap in the other. Scores for each pair of residues are read from a matrix. Those sequence pairs assumed to be the most similar are given the highest total scores. Although, as has already been stated, sequence-searching programs take amino acids or bases simply as characters, biochemical knowledge can be built into the matrix.

### Gap penalties

Clearly, aligning a residue or group of residues in one sequence with a null character (a 'gap') in another should be penalised. Since a single point mutation may introduce many more than one residue into a sequence, long gaps are usually penalised only slightly more than short ones. This is achieved by using two separate negative scores: a large penalty for introducing a gap and a much smaller one for extending an existing one.

### Scoring matrices

For comparisons between DNA sequences, the choice of scoring matrix is generally trivial. A high score is given for a match between bases and zero, or a negative score, to any mismatch. A match between A and 'not-A' is scored as a mismatch under all circumstances.

Comparisons at the protein level are much more complex. All algorithms comparing protein sequences give matches between amino acids thought of as 'similar' – such as leucine and isoleucine, or phenylalanine and tyrosine – intermediate scores between those of identical amino acids and those of amino acids with no similarity. Researchers have used different criteria to assign scores to each of the 210 possible pairs of amino acids.

### Genetic code schemes

One of the first schemes to be applied assumed that those changes that were most likely to occur were those that could arise from a point mutation in a single codon.<sup>11</sup> For example, it is



**Observed substitution, matrices, PAM, BLOSUM**

possible to mutate alanine into proline with a single point mutation, changing CCC into GCC. Changing leucine into isoleucine, however, takes a minimum of two codon changes. Using a matrix derived from the genetic code, an alignment of I with L scores lower than one of A with P.

**Chemical similarity schemes**

It is recognised that changing an amino acid into one of very different physicochemical type – for example, changing a large non-polar amino acid (such as phenylalanine) into a small polar one (such as serine) – is likely to disrupt the structure and function of the protein. Several workers including McLachlan<sup>12</sup> developed intuitive schemes to score changes between amino acids based on their chemical similarity. Feng *et al.*<sup>13</sup> later developed a scheme combining information about chemical properties and the genetic code.

**Observed substitution schemes**

Matrices based on chemical similarity and on the genetic code are still used in some applications. However, it is now recognised that for general database searches a third type of matrix – based on observed substitution schemes – tends to give more accurate matches for distantly related sequences. These matrices are derived by analysing how often one amino acid is seen to substitute for another in alignments of well-characterised protein families. One feature of this type of scheme is that identities are not all scored the same. A relatively common amino acid, such as alanine, will quite often occur at the same place in two aligned sequences by chance. This is much less likely to occur with a rare amino acid such as tryptophan. Therefore, tryptophan residues aligned together are typically given a higher score than similarly placed alanines.

In the 1970s, Margaret Dayhoff derived a set of matrices from observed substitution frequencies in a few widely

studied protein families. These matrices – the Dayhoff or PAM matrices<sup>14</sup> – are still in common use. Each element (defined as  $M_{A,B}$ ) in a PAM matrix reflects the probability of the amino acid in column A mutating into that in row B in a given length of evolutionary time, measured in Percentage of Acceptable point Mutations (PAM) per  $10^8$  years. Matrices representing greater evolutionary distances have larger PAM numbers. Using a matrix with a large PAM number will tend to find long, distantly related sequences while matrices with smaller PAM numbers will find shorter, more similar sequences. In many cases, the PAM250 matrix (Table 2) is seen as a good first choice. More recently, the PET91 matrix<sup>15</sup> has been calculated from a larger selection of homologous protein families using similar principles.

Dayhoff derived her matrices from a relatively small set of global alignments of very similar sequences. The BLOSUM (blocks substitution matrix) series of matrices were obtained using multiple local alignments of more distantly related sequences. Within each alignment set, sequences were clustered together into subfamilies of more than a given percentage sequence identity. A family of matrices can be obtained from substitution frequencies of aligned amino acids within these groups. Within such a family, individual matrices are distinguished by numbers indicating the level of sequence identity used in the original calculation. For example, the widely used BLOSUM62 matrix was derived from clusters of sequences of greater than 62 per cent identity. Matrices based on clusters of high sequence identity will find short, highly similar sequences. After a comparison of many observed substitution matrices,<sup>16</sup> Henikoff and Henikoff concluded that the BLOSUM62 matrix was the most effective overall.

It is likely that, as more and more protein structures are discovered, it will become possible to align distantly

**Table 2:** The PAM250 matrix – an example of a matrix derived from observed substitutions

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	2	0	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	0
B	0	2	-4	3	2	-5	0	1	-2	1	-3	-2	2	-1	1	-1	0	0	-2	-5	-3	2
C	-2	-4	12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-5
D	0	3	-5	4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	3
E	0	2	-5	3	4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	3
F	-4	-5	-4	-6	-5	9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7	-5
G	1	0	-3	1	0	-5	5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5	-1
H	-1	1	-3	1	1	-2	-2	6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	2
I	-1	-2	-2	-2	-2	1	-3	-2	5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	-2
K	-1	1	-5	0	0	-5	-2	0	-2	5	-3	0	1	-1	1	3	0	0	-2	-3	-4	0
L	-2	-3	-6	-4	-3	2	-4	-2	2	-3	6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	-3
M	-1	-2	-5	-3	-2	0	-3	-2	2	0	4	6	-2	-2	-1	0	-2	-1	2	-4	-2	-2
N	0	2	-4	2	1	-4	0	2	-2	1	-3	-2	2	-1	1	0	1	0	-2	-4	-2	1
P	1	-1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6	0	0	1	0	-1	-6	-5	0
Q	0	1	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4	1	-1	-1	-2	-5	-4	3
R	-2	-1	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6	0	-1	-2	2	-4	0
S	1	0	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2	1	-1	-2	-3	0
T	1	0	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3	0	-5	-3	-1
V	0	-2	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4	-6	-2	-2
W	-6	-5	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	0	-6
Y	-3	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10	-4
Z	0	2	-5	3	3	-5	-1	2	-2	0	-3	-2	1	0	3	0	0	-1	-2	-6	-4	3

The non-standard amino acid B is used to refer to (D or N); Z refers to (E or Q).

**Low complexity sequence**

related sequences much more accurately by including structural information. Analysis of these alignments is therefore likely to give better matrices than any of those derived from sequence data alone.

**Filtering**

**Filtering**  
The statistics used in database searching assume that the arrangement of bases or amino acids in unrelated sequences is essentially random. This is not the case in practice. In particular, some regions of DNA and protein sequences consist of long repeated runs of a single

residue or a pattern of residues. About 30 per cent of the human genome is known to consist of repetitive sequences of non-coding DNA. These features can also occur at the protein level, one example being polyaniline tracts. Repetitive sequences are not usually of interest to biologists, and close matches to these sequences may 'mask out' lower-scoring matches to homologous sequences. Therefore many implementations of popular database scanning programs allow users to filter out such sequences before running the search program. Typically,

**Algorithms**

nucleotides in low-complexity regions are replaced by the character N. Amino acids in such regions are filtered out by replacing them by the character X, using, for example, the program SEG.<sup>17</sup>

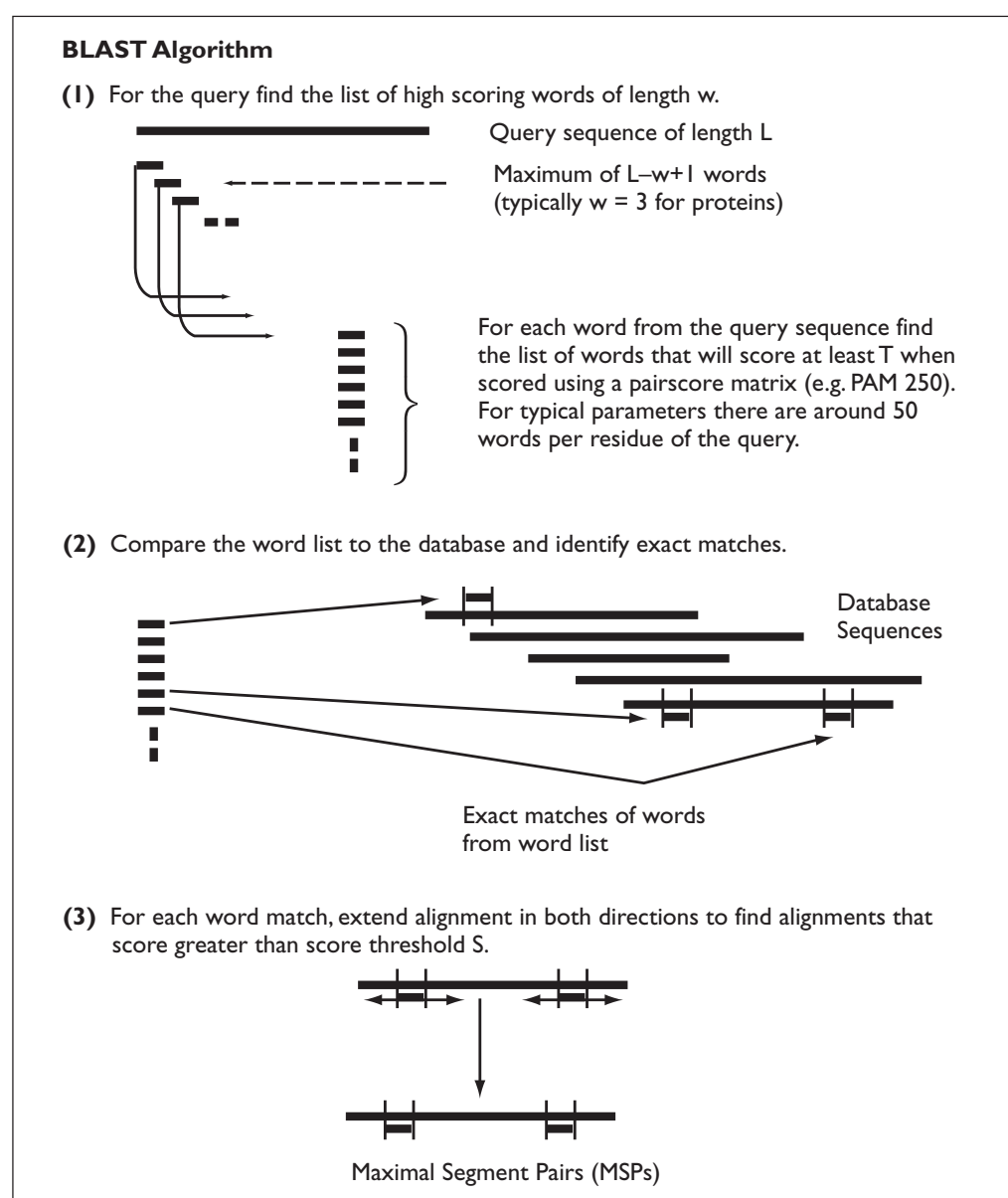
between a query sequence and a database. It is a 'heuristic' algorithm, implying that it uses a method that relies on 'guesses' to obtain approximately accurate results. Its speed is therefore achieved at the cost of some degree of precision. Later versions of BLAST (BLAST2 or Gapped BLAST), which allow gaps to be introduced into the alignments reported, are known to reflect biological relationships more accurately.

**ALGORITHMS**

**BLAST**

**BLAST**

BLAST<sup>2-4</sup> – Basic Local Alignment Search Tool – is a simple, but extremely fast, algorithm for finding the highest scoring locally optimal matches



**Figure 1:** Schematic illustration of the BLAST algorithm. Originally published in [http://barton.ebi.ac.uk/barton/papers/rev93\\_1/subsection3\\_7\\_6.htm/](http://barton.ebi.ac.uk/barton/papers/rev93_1/subsection3_7_6.htm/); reproduced with permission from Geoff Barton, European Bioinformatics Institute, Hinxton, UK. Taken from the HGMP-RC bioinformatics course material

The BLAST algorithm (Figure 1) operates as follows:

- A list of all possible short sequences (words,  $w$ ) in the query sequence that are of a given length and score greater than a cut-off value  $T$  using a particular scoring matrix is created. Each of these words will be similar to, but not necessarily identical to, a subsequence of the query sequence. The minimum level of similarity required is determined by  $T$ .
- The database is searched to retrieve every occurrence of each high-scoring word (the hit list).
- Each hit is extended to determine whether this match is part of a longer high-scoring sequence (scoring higher than a threshold  $S$ ). The BLAST results are presented as a sorted list of these high-scoring alignments (maximal sequence pairs or MSPs).

The statistics inherent in the algorithm, based on the work of Karlin and Altschul,<sup>18</sup> provide a direct estimate of the statistical significance of each match found. This is reported in the BLAST output as an  $E$ -value.

It is possible to alter several parameters in BLAST, to enable it to run faster or with higher precision. However, relatively inexperienced users will not need to change many of these parameters from their default values very often. The most often changed parameters are the scoring matrix (for searches conducted at the protein level); the gap creation and extension penalties; and the  $E$ -value taken as a cut-off. In some cases, particularly if the query sequence is quite short, sequences with relatively high  $E$ -values may be significant.

### FASTA

The FASTA algorithm for sequence database searching uses a fast procedure based on a method developed by Pearson and Lipman<sup>5,6</sup> that scans each

database sequence for those segments that are best matches to the test sequence. This is conceptually similar to finding the most significant diagonals in a 'dot-plot' of the two sequences (Figure 2). In the first step, the database is searched for those sequences that contain the largest number of aligned perfect matches to short sequences ('words') within the query sequence. For protein sequences, these significant matches are re-scored using one of the PAM or BLOSUM matrices. The top-scoring alignments are constructed by aligning all segments lying close to the diagonal containing the highest-scoring segments to the query sequence. The speed and sensitivity of the search are determined by the word length used in the first step. The shorter the words (also known as  $n$ -mers or  $k$ -tuples), the more sensitive the search will be and the longer it will take.

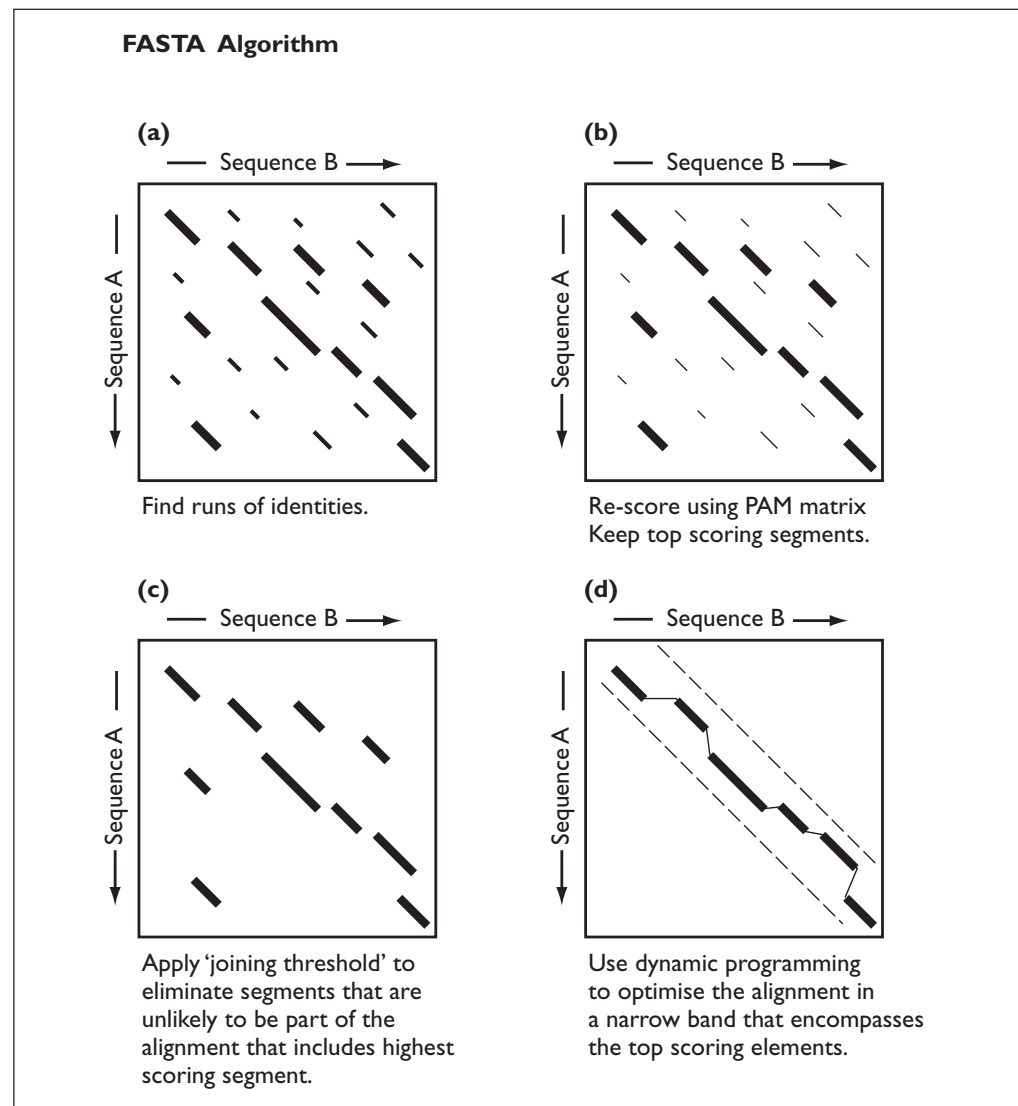
Besides choosing a scoring matrix and gap penalties, the FASTA user may wish to set the word length with the parameter  $ktup$ . In protein database searching, setting  $ktup$  to 1 will run a slow, sensitive search; setting it to 2 will run a fast search that is not much more sensitive than BLAST. Word lengths used in DNA sequence searches are generally longer, but the maximum value of  $ktup$  in general use is six.

BLAST and FASTA allow any combination of DNA and protein test sequences to be used to search any combination of protein and DNA databases. Some implementations select the program to be used based on the type of query sequence and the database selected; others expect the user to choose it explicitly. As an example, the names sometimes given to the different programs in the BLAST family are given in Table 3.

### PSI-BLAST

PSI-BLAST, or position-specific iterated BLAST,<sup>4</sup> is an extension to the BLAST algorithm which is an extremely sensitive method of determining

### FASTA



**Figure 2:** Schematic illustration of the FASTA algorithm. Originally published in [http://barton.ebi.ac.uk/barton/papers/rev93\\_1/subsection3\\_7\\_7.html/](http://barton.ebi.ac.uk/barton/papers/rev93_1/subsection3_7_7.html/); reproduced with permission from Geoff Barton, European Bioinformatics Institute, Hinxton, UK. Taken from the HGMP-RC bioinformatics course material

**Table 3:** Names of programs in the BLAST family. Similar options are available with FASTA

Program name	Function
<b>Blastp</b>	Compares an amino acid sequence against a protein database
<b>Blastn</b>	Compares a nucleic acid sequence against a nucleic acid database
<b>Blastx</b>	Translates a nucleic acid query sequence in all six reading frames and compares each conceptual translation against a protein database
<b>Tblastn</b>	Compares a protein query sequence against a nucleic acid database where each sequence has been dynamically translated in all six reading frames
<b>Tblastx</b>	Compares the six-frame translations of a nucleic acid sequence against the six-frame translations of a nucleic acid database (the comparison being at the protein level)

**PSI-BLAST****Profile****Hidden Markov model**

protein sequence homologies. The procedure starts with a simple BLAST search with a single protein sequence. The resulting 'hits' are extracted, aligned and formed into a 'profile' containing information from all sequences in the family. The next stage is another BLAST search of the database using the profile instead of a single test sequence. This procedure is repeated until one iteration produces no significant new matches.

Although PSI-BLAST is extremely sensitive, and will often discover more remote homologies than either BLAST or FASTA, it needs to be treated with care. If a single unrelated sequence is included in the alignment at one stage it will form part of the test sequence set throughout the rest of the run, with unpredictable results. Eddy<sup>19</sup> described a case where a PSI-BLAST run wrongly classified an uncharacterised nematode sequence believed to be a G protein-coupled receptor as a ribosomal L11 protein, based on the inclusion of a borderline match in the profile. It is clear that the output from each iteration in a PSI-BLAST run should be scrutinised carefully.

**Other programs**

Some other programs for database searching, which are used less frequently, can be very useful in some circumstances. Most of these are procedures for searching a database with a single sequence that are more rigorous, but slower, than either BLAST or FASTA. They include Blitz and SSEARCH,<sup>20</sup> which implements the Smith–Waterman algorithm often used in pairwise sequence alignment calculations to perform a rigorous comparison of the test sequence with each sequence in the database.

Some of the most interesting recent developments in the field of sequence analysis algorithms involve the use of a class of statistical methods termed hidden Markov models (HMMs) to describe alignments of related sequences. Hidden Markov models have been used to model other 'linear' systems, and applied to problems such as speech recognition, for many years; a full description of HMM theory is far beyond the scope of this review. Methods such as PSI-BLAST<sup>4</sup> use 'profiles' derived from many aligned members of sequence families to

**Table 4:** Useful URLs for sequence database searching

<b>Sequence databases</b>	
EMBL database at EBI	<a href="http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html">http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html</a>
GenBank database at NCBI	<a href="http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html">http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html</a>
Swiss-Prot and TrEMBL	<a href="http://www.expasy.ch/">http://www.expasy.ch/</a>
The Protein DataBank*	<a href="http://www.rcsb.org/">http://www.rcsb.org/</a>
Whole genome databases	<a href="http://www.tigr.org/tdb/mdb/mdb.html">http://www.tigr.org/tdb/mdb/mdb.html</a>
<b>Search algorithms</b>	
BLAST website	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>
FASTA website	<a href="http://www2.ebi.ac.uk/fasta3/">http://www2.ebi.ac.uk/fasta3/</a>
PSI-Blast server	<a href="http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast/">http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast/</a>
HGMP Resource Centre: Blast, FASTA†	<a href="http://www.hgmp.mrc.ac.uk/">http://www.hgmp.mrc.ac.uk/</a>

\* Formerly <http://www.pdb.bnl.gov>.

† General bioinformatics service; registration is necessary (free for academics).

EBI, European Bioinformatics Institute; NCBI, National Center for Biotechnology Information.

identify more distantly related sequences. In simple terms, an HMM profile is a rigorous description of a probability distribution over an infinite number of possible sequences.<sup>21</sup> Programs such as HMMer (S. R. Eddy, unpublished) that search databases with profiles derived in this way are accurate and sensitive, but can be extremely slow.

In conclusion, given the impressive growth of the sequence databases (Table 4), and the growing importance of this subject, most molecular biologists will soon need to know at least the basic principles of database sequence searching. Although there is a wide range of algorithms to choose from, the fastest and simplest – the BLAST series of algorithms – should be sufficient for many purposes. One should not, however, have blind faith in the results of any algorithm, however rigorous. Database searches are computer simulations, and the results gained are dependent on the assumptions made in planning the search. It will always be important to check database search results against biological intuition and, where possible, with ‘wet’ biology.

### References

1. Doolittle, R. F. (1997), ‘Some reflections on the early days of sequence searching’, *J. Mol. Med.*, Vol. 75, pp. 239–241.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990), ‘Basic local alignment search tool’, *J. Mol. Biol.*, Vol. 215, pp. 403–410.
3. Altschul, S. F. and Gish, W. (1996), ‘Local alignment statistics’, *Methods Enzymol.*, Vol. 266, pp. 460–480.
4. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997), ‘Gapped BLAST and PSI-BLAST: a new generation of protein database search programs’, *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.
5. Pearson, W. R. and Lipman, D. J. (1988), ‘Improved tools for biological sequence comparison’, *Proc. Nat. Acad. Sci. USA*, Vol. 85, pp. 2444–2448.
6. Lipman, D. J. and Pearson, W. R. (1985), ‘Rapid and sensitive protein similarity searches’, *Science*, Vol. 227, pp. 1435–1441.
7. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. and Wheeler, D. L. (1999), ‘GenBank’, *Nucleic Acids Res.*, Vol. 27, pp. 12–17.
8. Bairoch, A. and Apweiler, R. (1999), ‘The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999’, *Nucleic Acids Res.*, Vol. 27, pp. 49–54.
9. Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., Abola, E. E. (1998), ‘Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules’, *Acta Crystallogr.*, Vol. D54, pp. 1078–1084. Now at <http://www.rcsb.org/pdb/>.
10. Johnson, G., Kabat, E. A. and Wu, T. T. (1996), ‘Kabat Database of Sequences of Proteins of Immunological Interest’, in ‘Weir’s Handbook of Experimental Immunology I. Immunochemistry and Molecular Immunology, Fifth Edn’, Herzenberg, L. A., Weir, W. M., Herzenberg, L. A. and Blackwell, C., Eds, Blackwell Science Inc, Cambridge, Mass. Chapter 6, pp. 6.1–6.21.
11. Fitch, W. M. (1966), ‘An improved method of testing for evolutionary homology’, *J. Mol. Biol.*, Vol. 16, pp. 9–16.
12. McLachlan, A. D. (1972), ‘Repeating sequences and gene duplication in proteins’, *J. Mol. Biol.*, Vol. 64, pp. 417–437.
13. Feng, D. F., Johnson, M. S. and Doolittle, R. F. (1985), ‘Aligning amino acid sequences: comparison of commonly used methods’, *J. Mol. Evol.*, Vol. 21, pp. 112–125.
14. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978), ‘A model of evolutionary change in proteins: Matrices for detecting distant relationships’, in Dayhoff, M. O., Ed., ‘Atlas of Protein Sequence and Structure’, Vol. 5, National Biomedical Research Foundation, Washington DC, pp. 345–358.
15. Henikoff, S. and Henikoff, J. G. (1992), ‘Amino acid substitution matrices from protein blocks’, *Proc. Natl Acad. Sci. USA.*, Vol. 89, pp. 10915–10919.
16. Henikoff, S. and Henikoff, J. G. (1993), ‘Nonglobular domains in protein sequences: automated segmentation using RT complexity measures’, *Proteins*, Vol. 17, pp. 49–61.
17. Wootton, J. C. (1994), ‘Nonglobular domains in protein sequences: automated segmentation using RT complexity measures’, *Comput. Chem.*, Vol. 18, pp. 269–285.
18. Karlin, S. and Altschul, S. F. (1990), ‘Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes’, *Proc. Natl Acad. Sci.*, Vol. 87, pp. 2264–2268.



19. Eddy, S. R. (1998), 'Trends Guide to Bioinformatics', Elsevier Trends Journals, Cambridge, pp. 15–18.
20. Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.*, Vol. 147, pp. 195–197.
21. Eddy, S. R. (1996), 'Hidden Markov models', *Curr. Opin. Struct. Biol.*, Vol. 6, pp. 361–365.

## APPENDIX 1

### *Practical Guidelines for Database Searching: A Short Summary*

- Search against an up-to-date database that is relevant to your query.
- It is probably worth starting any search strategy by using a fast algorithm such as BLAST.
- Use the latest version of your chosen algorithm, eg Gapped BLAST.
- Work at the protein level if appropriate.
- Filter out any low-complexity regions.
- Start with a general scoring matrix such as PAM250 or BLOSUM62:
  - if few matches are found, try higher-scoring matrices: PAM400 or BLOSUM30;
  - if the members of the sequence family are very similar, try realigning with lower scoring matrices: PAM40 or BLOSUM80;
  - if few appropriate matches are still found, try a more precise method.
- Check all results; if your biochemical intuition tells you there is something wrong with the search results, there probably is.
- Repeat searches often, particularly if you are working in a fast-moving subject area.

