parameters[3,10,13] of the scoring system employed, and is therefore more informative than a raw score for describing the quality of an alignment.

Masking regions of restricted composition

Many DNA and protein sequences contain regions of highly restricted nucleic acid and amino acid composition and regions of short elements repeated many times[15]. The standard alignment models and scoring systems were not designed to capture the evolutionary processes that led to these **low-complexity regions**. As a result, two sequences containing compositionally biased regions can receive a very high similarity score that reflects this bias alone. For many purposes, these regions are uninteresting and can obscure other important similarities. Therefore, programs that filter low-complexity regions from query or database sequences will often turn a useless database search into a valuable one[15].

Multiple sequences

Global and local pairwise sequence comparison and alignment can be generalized to multiple sequences. From multiple alignments, **profiles** [related to hidden Markov models (**HMMs**)] can be abstracted and these can greatly enhance the sensitivity of database search methods to evolutionarily distant and subtle sequence relationships[11]. As discussed by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21, this area is increasingly becoming the focus of algorithm and database development for biological sequence comparison.

Dedication

This article is dedicated to Dr Bruce W. Erickson, friend and mentor.

References

1 Dayhoff, M.O. *et al.* (1978) in *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–352, National Biomedical Research Foundation
2 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
3 Altschul, S.F. (1991) *J. Mol. Biol.* 219, 555–565
4 Pearson, W.R. (1995) *Protein Sci.* 4, 1145–1160
5 Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443–453
6 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
7 Gotoh, O. (1982) *J. Mol. Biol.* 162, 705–708
8 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
9 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
10 Altschul, S.F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
11 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
12 Fitch, W.M. (1983) *J. Mol. Biol.* 163, 171–176
13 Karlin, S. and Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
14 Dembo, A. *et al.* (1994) *Ann. Probab.* 22, 2022–2039
15 Wootton, J.C. and Federhen, S. (1993) *Comput. Chem.* 17, 149–163

Practical database searching

Sequence comparisons need to be performed as carefully as wet-lab procedures, in terms of both experimental design and interpretation. The basic requirements of database searching, the factors that can affect the search results and, finally, how to interpret the results are discussed.

Steven E. Brenner
Dept of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA.
brenner@hyper.stanford.edu

More sequences have been putatively characterized by database searches than by any other single technology. For good reason: programs like **BLAST** are fast and reliable. However, sequence comparison procedures should be treated as experiments analogous to standard laboratory procedures. Their use deserves the same care both in the design of the experiment and in the interpretation of results.

The database search experiment

Design of a BLAST database search requires consideration of what information is to be gained about the query sequence of interest. The main constraint is that database searching can only reveal similarity. However, from this similarity, homology (i.e. evolutionary relationship) can be inferred and, from that, one might be able to infer function. Although the former inference is now reliable

Box 1. Database searching: basic considerations

- Think about every step
- Search a large current database
- Compare as protein rather than DNA
- Filter query for low-complexity regions
- Interpret scores with *E* values
- Recognize that most homologs are not found by pairwise sequence comparison
- Consider slower and more powerful methods, but use iterative programs with great caution

for carefully performed sequence comparison, the second is still fraught with challenges. Box 1 provides some guidelines for performing reliable and sensitive database searches.

Planning a good experiment requires an understanding of the method being applied. Fundamentally, database searches are a simple operation: a query sequence is locally aligned with each of the sequences (called targets) in a database. Most programs, such as BLAST (Ref. 1) and **FASTA** (Ref. 2), use **heuristics** to speed up the alignment procedure, while the **Smith–Waterman algorithm**[3] (implemented, for example, in **SSEARCH**) rigorously compares the query sequence with each target in the database.

A score is computed from each alignment, and the query–target pairs with the best scores are then reported to the user. Typically, statistics are used to help improve the interpretation of these scores. A more detailed description of the process can be found in the article by Stephen Altschul on pp. 7–9. Although BLAST is the most widely used tool for sequence comparison, many other programs can help identify, confirm and interpret distant evolutionary relationships.

Databases, programs and comparison types

Formulation of the experiment begins with a decision about what types of sequences to compare: DNA, protein or DNA as protein. If the sequence under consideration is a protein or codes for a protein, then the search should probably take place at the protein level, because proteins allow one to detect far more distant homology than does DNA[2,4]. For example, in DNA comparisons, there is noise from the rapidly mutated third-base position in each codon and from comparisons of noncoding frames (although this latter issue still arises in DNA-as-protein searches). In addition, amino acids have chemical characteristics that allow degrees of similarity to be assessed rather than simple recognition of identity or non-identity. For these reasons, DNA versus DNA comparison (using the blastn program) is typically only used to find identical regions of sequence in a database. One would carry out such a search to discover whether the gene has been previously sequenced and to determine where it is expressed or where splice junctions occur. In short, protein-level searches are valuable for detecting evolutionarily related genes, while DNA searches are best for locating nearly identical regions of sequence.

Next, it is necessary to select a database to search against. For homology searches, the most commonly searched database on the NCBI (National Center for Biotechnology Information) website is the **nr database**. The nr protein database combines data from several sources, removes the redundant identical sequences and yields a collection with nearly all known proteins. The NCBI nr database is frequently updated in order to incorporate as many sequences as possible. Obviously, a search will not identify a sequence that has not been included in the database and, as databases are growing so rapidly, it is essential to use a current database. Several specialized databases are also available, each of which is a subset of the nr database. ***E*-value** statistics (discussed below) are affected by database size, so, if you are interested in searching for proteins of known structure, it is best to just search the smaller **pdb database**.

One might also wish to search DNA databases at the protein level. Programs can do so automatically by first translating the DNA in all six reading frames and then making comparisons with each of these conceptual translations. The nr DNA database, which contains most known DNA sequences except **GSS**s, **EST**s, **STS**s and **HTGS**s, is useful to search when hunting new genes; the identified genes in this database would already be in the protein nr database. Searches against the GSS, EST, STS and HTGS databases can find new homologous genes and are especially useful for learning about expression data or genome map location.

Because of the different combinations of queries and database types, there are several variants of BLAST (see Table 1). Note that it is desirable to use the newest versions of BLAST, which support **gapped alignments** (see the article by Stephen Altschul on pp. 7–9). The older versions are slower, detect fewer homologs and have problems with some statistics. The programs can be run over the World Wide Web (WWW) and can be downloaded from an **ftp** site to run locally. Another option is to use the FASTA package[2]. The FASTA program can be slower but more effective than BLAST. The package also contains SSEARCH, an implementation of the rigorous Smith–Waterman algorithm, which is slow but the most sensitive. As described in the article by Sean Eddy on pp. 15–18, iterative programs such as **PSI-BLAST** require extreme care in their operation because they can provide very misleading results; however, they have the potential to find more homologs than purely pairwise methods.

Filtering

The statistics for database searches assume that unrelated sequences will look essentially random with respect to each other. However, certain patterns in sequences violate this rule. The most common exceptions are long runs of a small number of different residues (such as a polyalanine tract). Such regions of sequence could spuriously obtain extremely high match scores. For this reason, the NCBI BLAST server will automatically remove such sections in proteins (replacing them with an X) using the **SEG** program[5] if 'default **filtering**' is selected. DNA sequences will be similarly **masked** by **DUST**.

Table 1. BLAST variants for different searches[a]

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Seek identical DNA sequences and splicing patterns
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Analyze new DNA to find genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Search for genes in unannotated DNA
tblastx	DNA	DNA	Protein level	Discover gene structure

[a]Similar variant programs are available for FASTA. Protein-level searches of DNA sequences are performed by comparing translations of all six reading frames.

Although these programs automatically remove the majority of problematic matches, some problems invariably slip through; moreover, valid hits might be missed if part of the sequence is masked. Therefore, it might be helpful to try using different masking parameters.

Other sorts of filtering are also often desirable. For example, **iterative searches** are prone to contamination by regions of proteins that resemble coiled coils or transmembrane helices. The problem is that a protein that is similar only in these general characteristics might match initially. The profile then emphasizes these inappropriate characteristics, eventually causing many spurious hits. Heavily cysteine-rich proteins can also obtain anomalous high scores. Especially if these characteristics are not filtered, it is necessary to review the alignment results carefully to ensure that they have not led to incorrect matches.

Alignment, algorithmic and output parameters

Three other sets of parameters also affect search results, but they rarely require careful consideration by most users. First, the matrix and gap parameters determine how similarity between two sequences is determined. When two residues in a protein are aligned, programs use the matrix to determine whether the amino acids are similar (and thus receive a positive score) or very different. The default matrix for BLAST is called BLOSUM62 (Ref. 6), and the programs will not currently operate reliably with other matrices. The gap parameters determine how much an alignment is penalized for having gaps: the existence parameter is a fixed cost for having a gap and the per-position cost is a cost dependent upon the length (i.e. the number of residues). Typically, there is a large cost associated with introducing a gap and a small additional cost such that longer gaps are worse. It is rarely very beneficial to change these from their defaults.

The second set of parameters determines the heuristics that BLAST uses. By altering these numbers, it is possible to make the program run slower and be more sensitive, or to run faster at the cost of missing more homologs. The complexity of these parameters in BLAST precludes extensive description here. Currently, it is very rare for users to alter these options from the defaults. The FASTA program has one such parameter, called **ktup**, that a user will often want to set. Searches with ktup = 1 are slower, but more sensitive, than BLAST; ktup = 2 is fast, but less effective.

A third set of parameters regulates how many results are reported. By default, the programs will report only matches with an ***E* value** (described below) up to 10. The total number of matches is limited to the best 500, and detailed information with the alignment is provided for up to 100 pairs. To retrieve more matches, these numbers can be increased.

Interpretation of results

Interpretation of the results of a sequence database search involves first evaluating the matches to determine whether they are significant and therefore imply homology. The most effective way of doing this is through use of statistical scores or *E* values. The *E* values are more useful than the **raw** or **bit scores**, and they are far more powerful than percentage identity (which is best not even considered unless the identity is very high)[7]. Fortunately, the *E* values from FASTA, SSEARCH and NCBI gapped BLAST seem to be accurate and are therefore easy to interpret (see Ref. 7).

The *E* value of a match should measure the expected number of sequences in the database that would achieve a given score by chance. Therefore, in the average database search, one expects to find ten random matches with *E* values below 10; obviously, such matches are not significant. However, lacking better matches, sequences with these scores might provide hints of function or suggest new experiments. Scores below 0.01 would occur by chance only very rarely and are therefore likely to indicate homology, unless biased in some way. Scores of near 1e−50 (1×10^{-50}) are now seen frequently, and these offer extremely high confidence that the query protein is evolutionarily related to the matched target in the database.

Inferring function from the homologous matched sequences is a problematic process. If the score is extremely good and the alignment covers the whole of both proteins, then there is a good chance that they will share the same or a related function. However, it is dangerous to place too much trust in the query having the same function as the matched protein; functions do diverge, and organismal or cellular roles can alter even when biochemical function is unchanged. Moreover, a significant fraction of functional annotations in databases are wrong, so one needs to be careful. There are other complexities; for example, if only a portion of the proteins align, they might share a domain that only contributes one aspect of the overall function. It is often the case that all of the highest-scoring hits align to one region of the query, and matches to other regions need to be sought much lower

BLAST Web site
http://www.ncbi.nlm.nih.gov/BLAST/

BLAST FTP site
ftp://ncbi.nlm.nih.gov/blast/

FASTA at EBI
http://www2.ebi.ac.uk/fasta3/

FASTA FTP site
ftp://ftp.virginia.edu/pub/fasta

Sequence search site
http://sss.stanford.edu/sss/

in the score ranking. For this reason, it is necessary to consider carefully the overlap between the query and each of the targets.

Database search methods are also limited because most homologous sequences have diverged too far to be detected by pairwise sequence comparison methods[7]. Thus, failure to find a significant match does not indicate that no homologs exist in the database; rather, it suggests that either more-powerful computational methods (such as those described by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21) or experiments would be necessary to locate them.

Conclusion

One should neither have excessive faith in the results of a BLAST run nor blithely disregard them. The BLAST programs are well-tested and reliable indicators of sequence similarity, and their underlying principles are straightforward. Complexities added by the fast algorithms typically need not be carefully considered, because the program and its parameters have been optimized for hundreds of thousands of daily runs. If one is careful about posing the database search experiment and interprets the results with care, sequence comparison methods can be trusted to provide an incomparable wealth of biological information rapidly and easily.

References

1 Altshul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
2 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
3 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
4 States, D.J. *et al.* (1991) *Methods* 3, 66–70
5 Wootton, J.C. (1994) *Comput. Chem.* 18, 269–285
6 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
7 Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078

Computational genefinding

A major challenge in the analysis of genomic DNA sequence is to find the functional sites that encode elements responsible for gene structure, regulation and transcription. A variety of computational tools can help to isolate the 'signal' from the 'noise'.

David Haussler
Computer Science Dept, University of California, Santa Cruz, CA 95064, USA.
haussler@cse.ucsc.edu

Expanded version of this article
http://www.cse.ucsc.edu/~haussler/pubs.html

Computational methodology for finding genes and other functional sites in genomic DNA has evolved significantly over the past 20 years (for reviews, see Refs 1–3). The genomic elements that researchers seek include splice sites, start and stop codons, branch points, promoters and terminators of transcription, polyadenylation sites, ribosomal binding sites, topoisomerase II binding sites, topoisomerase I cleavage sites and various transcription factor binding sites[4]. Local sites such as these are called **signals**, and methods for detecting them can be called signal sensors. In contrast, extended and variable-length regions, such as exons and introns, are called **contents** and are recognized by methods that can be called content sensors[5].

Signal sensors

The most basic signal sensor is a simple consensus sequence, or an expression that describes a consensus sequence together with allowable variations. More sensitive sensors can be designed using **weight matrices** in place of the consensus, in