BLAST Algorithm

Chapter · September 2005

DOI: 10.1038/npg.els.0005253

1 author:

Stephen F Altschul
National Institutes of Health
107 PUBLICATIONS **222,400** CITATIONS

SEE PROFILE

BLAST Algorithm

Advanced article

Stephen F Altschul, *National Center for Biotechnology Information, Bethesda, Maryland, USA*

doi: 10.1038/npg.els.0005253

BLAST is an acronym for basic local alignment search tool; the BLAST family of database search programs takes as input a query DNA or protein sequence, and search DNA or protein sequence databases for similarities that may indicate homology. The programs implement variations of the BLAST algorithm, which is a heuristic method for rapidly finding local alignments with scores sufficiently high to be statistically significant.

Deoxyribonucleic acid (DNA) or protein sequences are aligned with one another for a wide variety of reasons, for example, to infer homology, to predict protein structure and function, to reconstruct phylogeny and to locate exons and introns. The type of sequence alignment that makes the fewest prior assumptions is local alignment, which seeks similar segments of unspecified length from the two sequences being compared. Local sequence similarity is generally measured with the aid of a nucleotide or amino acid substitution matrix and associated gap costs. With sequence similarity so defined, a rigorous method for finding optimal local alignments is provided by the Smith–Waterman algorithm, which requires time proportional to the product of the lengths of the sequences it compares. However, using this algorithm, a similarity search of a large nucleic acid sequence database or protein database with a single query sequence of moderate length may require an hour on a modern work station. Accordingly, rapid heuristic algorithms such as FASTA and basic local alignment search tool (BLAST) have been developed that can perform these searches up to two orders of magnitude faster than the Smith–Waterman algorithm, but at the cost of possibly missing an occasional similarity, that is, local alignment, of interest.

BLAST Statistics

Local alignments of the kind sought by BLAST are evaluated by alignment scores, defined as the sum of substitution scores for aligned pairs of letters, and gap scores for strings of letters in one sequence aligned with null characters introduced into the other. Whether an alignment score is great enough to be of interest depends on what scores are expected to arise purely by chance. An analytic theory exists only for local alignments that may not contain gaps. Assuming that 'random' DNA or protein sequences are strings of nucleotides or amino acids chosen independently, with fixed background residue frequencies, the number of distinct local alignments with score at least S expected to arise from the comparison of two sequences of lengths m and n is well approximated by the formula

$$E = Kmn\,\mathrm{e}^{-\lambda S}$$

where K and λ are calculable parameters (Karlin and Altschul, 1990). No such theory has been established for gapped local alignments, but computational experiments strongly suggest that the formula still applies in this more general case (Altschul and Gish, 1996). However, in the gapped case, the statistical parameters λ and K can no longer be derived analytically but must rather be estimated by random simulation.

The BLAST programs convert raw local alignment scores to 'E-values' using the above equation, m being the length in residues of the query sequence, and n the length of the database to which it is compared. The original BLAST programs (Altschul *et al.*, 1990) sought only ungapped local alignments and therefore could calculate λ and K analytically. The gapped BLAST programs (Altschul and Gish, 1996; Altschul *et al.*, 1997) allow users to choose among a set of substitution and gap costs for which the statistical parameters have been preestimated. Only local alignments with an E-value lower than a set threshold, usually 10 by default, are reported.

Basic Algorithm

Protein similarities of great biologic significance may be quite diffuse, sometimes involving long alignments that nowhere contain more than two adjacent matching pairs of amino acids. No rigorous method substantially faster than the Smith–Waterman algorithm is known for locating optimal local alignments. On the other hand, computer science has very rapid

techniques for identifying perfectly matching substrings of two input sequences. The heuristic BLAST algorithm uses these techniques to locate, with strong probability, optimal local alignments, even when they represent relatively diffuse similarities.

Let a 'word' be a set-length string of adjacent letters within a sequence. Given two sequences and a threshold score T, let a 'hit' be a pair of words, one from each sequence, whose aligned score is at least T. The central idea behind the BLAST algorithm is that for an appropriately chosen T, any local alignment between the query and a database sequence with an E-value worth reporting is highly likely to contain an aligned pair of words constituting a hit (Altschul *et al*., 1990). Therefore, a rapid search for hits involving query and database sequences can be used to locate candidate local alignments for reporting.

For the purpose of illustration, consider the original ungapped BLAST algorithm (Altschul *et al*., 1990) applied to protein sequence comparison using word length 3, and the threshold $T = 11$ for hits scored using the BLOSUM-62 amino acid substitution matrix (Henikoff and Henikoff, 1992). Omitting most details, BLAST follows the following steps when supplied with a query sequence and database.

Firstly, construct a table of all possible words that would form a hit when aligned with some word in the query sequence, and tag each word in the table with the locations of all query sequence words that generated it. For example, if the word 'MVD' appeared somewhere in the query, it would generate the 11 words shown in **Table 1**, each with a record of the location of 'MVD' in the query.

Secondly, scan the database for any words that appear in the table. Each such database word then constitutes a hit in conjunction with one or more known query sequence words.

Table 1 High-scoring words that align to 'MVD'

Word	Score
MVD	15
MID	14
L**VD**	12
MLD	12
MMD	12
I**VD**	11
V**VD**	11
MAD	11
MTD	11
MVE	11
LI**D**	11

Alignment scores are computed using the BLOSUM-62 amino acid substitution matrix (Henikoff and Henikoff, 1992). Amino acids matching those to which they are aligned in 'MVD' are shown in bold.

Thirdly, consider each hit a nascent alignment and extend it in each direction by aligning additional pairs of amino acids from the query and database sequences. Terminate an extension if the alignment score drops more than a fixed amount X below the best score yet found seeded by this hit. Report the optimal alignment for this hit if its score has a sufficiently low E-value.

BLAST achieves its speed in large measure by scanning the database only for *exact* matches to words in the table created in the first step above. However, because it generates this table by expanding each query word into a set of similar words, BLAST is able to detect fairly diffuse similarities. For example, six of the 11 hits generated by 'MVD' do not contain two adjacent pairs of matching amino acids (**Table 1**).

A key feature of the BLAST algorithm is its tunable trade-off between speed and the probability of missing relevant alignments. If the hit threshold score is raised from 11 to 12, the number of words generated by 'MVD', for example, falls from 11 to 5 (**Table 1**). This results in many fewer alignment extensions and a corresponding decrease in execution time, but it also renders it more likely that an alignment of interest will contain no hit and therefore be overlooked. Gapped versions of BLAST have additional conditions on and procedures for generating gapped alignments (Altschul and Gish, 1996; Altschul *et al*., 1997), which will not be discussed here.

BLAST Variants

Variations on the BLAST algorithm have been implemented for use in a variety of contexts.

Protein sequence comparison: BLASTP

BLASTP is used to compare a protein query sequence to a database of protein sequences; recent versions allow gapped alignments (Altschul and Gish, 1996; Altschul *et al*., 1997).

DNA sequence comparison: BLASTN and MEGABLAST

BLASTN is used to compare a DNA query sequence to a database of DNA sequences. In lieu of the score-based hit definition used for protein comparisons, hits are defined as runs of matching pairs of nucleic acids of a specified minimum length. MEGABLAST (Zhang *et al*., 2000) is a faster variant on BLASTN best used for very long query sequences and for finding alignments of very closely related sequences.

Translating comparisons: BLASTX, TBLASTN and TBLASTX

BLASTX (Gish and States, 1993) compares a DNA query sequence to a protein database. The query sequence is conceptually translated into protein in all six possible reading frames, and the comparisons are performed using an amino acid substitution matrix. TBLASTN compares a protein query sequence to a DNA sequence database. Database sequences are conceptually translated in all six reading frames, and the comparisons are performed at the protein level. TBLASTX compares a DNA query sequence to a DNA database, but at the protein level. Both query and database sequences are translated into protein in all possible reading frames, resulting in 36 comparisons for each database sequence.

Profile searches: PSI-BLAST, IMPALA and RPS-BLAST

PSI-BLAST (Altschul *et al*., 1997), an acronym for 'position-specific iterated BLAST', compares a protein query sequence to a protein database in an iterative manner. After each round of searching, any statistically significant alignments are combined into a multiple alignment, from which a position-specific score matrix or profile is abstracted. This profile is compared to the database in the next round of searching. PSI-BLAST has the potential to detect much more subtle sequence relationships than BLAST. IMPALA (Schäffer *et al*., 1999) and RPS-BLAST compare a protein query sequence to a database of PSI-BLAST generated profiles; the former uses the Smith–Waterman algorithm for the comparison and the latter a variant of the BLAST algorithm.

Pattern-based searches: PHI-BLAST

PHI-BLAST (Zhang *et al*., 1998), an acronym for 'pattern-hit initiated BLAST', compares a protein query sequence to a protein database but requires any alignments found to contain a user-specified pattern. This pattern serves in place of a hit table to generate candidate alignments for extension and evaluation.

Implementations and Servers

A number of BLAST servers are freely available on the internet (see Web Links). These servers employ different implementations of the BLAST programs. Executable code is available from both sites, and source code is available from the National Center for Biotechnology Information website (Wheeler *et al*., 2001).

See also

FASTA Algorithm
Protein Homology Modeling
Sequence Similarity
Similarity Search
Smith–Waterman Algorithm

References

Altschul SF and Gish W (1996) Local alignment statistics. *Methods in Enzymology* **266**: 460–480.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Altschul SF, Madden TL, Schäffer AA, *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

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

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 10 915–10 919.

Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 2264–2268.

Schäffer AA, Wolf YI, Ponting CP, *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**: 1000–1011.

Wheeler DL, Church DM, Lash AE, *et al*. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **29**: 11–16.

Zhang Z, Schäffer AA, Miller W, *et al*. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research* **26**: 3986–3990.

Zhang Z, Schwartz S, Wagner L and Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.

Further Reading

Altschul SF and Koonin EV (1998) Iterated profile searches with PSI-BLAST: a tool for discovery in protein databases. *Trends in Biochemical Sciences* **23**: 444–447.

Altschul SF, Boguski MS, Gish W and Wootton JC (1994) Issues in searching molecular sequence databases. *Nature Genetics* **6**: 119–129.

Altschul SF, Bundschuh R, Olsen R and Hwa T (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research* **29**: 351–361.

Baxevanis AD and Ouellette BFF (2001) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd edn. New York: John-Wiley.

Dembo A, Karlin S and Zeitouni O (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Annals of Probability* **22**: 2022–2039.

Ewens WJ and Grant GR (2001) *Statistical Methods in Bioinformatics*. New York: Springer-Verlag.

Schäffer AA, Aravind L, Madden TL, *et al*. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**: 2994–3005.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

Smith TF, Waterman MS and Burks C (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13**: 645–656.

Wilson AC, Carlson SS and White TJ (1977) Biochemical evolution. *Annual Review of Biochemistry* **46**: 573–639.

Web Links

National Center for Biotechnology Information BLAST server
http:/www.ncbi.nlm.nih.gov/BLAST/

Washington University BLAST server
http:/blast.wustl.edu/