Heuristic string comparison

Dynamic programming methods for sequence alignment give the highest quality results.

However, quadratic $O(nm)$ algorithms are only feasible for comparing two modest sized sequences.

Comparing the human genome against mouse with a quadratic algorithm ($3,000,000,000^2$ operations) at a billion operations per second equals 285 years!

Comparing your target sequence against the entire database is hopeless, even with special purpose hardware.

Instead, heuristic algorithms such as BLAST and FASTA are used to make an initial scan of the database to find a small number of hits.

Smith-Waterman can then be used to find the optimal alignment to display.

FASTA

FASTA is a heuristic string alignment program which is based upon finding short exact matches (k-mers) between the query sequence and the database.

The trick is to choose k large enough that there are relatively few hits, but small enough that we are likely to have an exact k-mer match between related sequences.

Recommended values of k are 2 for protein sequences and 6 for DNA sequences.

Note that there are at most $n - k + 1$ distinct k-mers in a sequence of length n.

A *hash table* of all k-mers in the database can be built once and used repeatedly for efficiently looking up the k-mers in query strings.

A dynamic programming-like algorithm is used to align the k-mer hits between query and database, but the problem is much smaller since there are far fewer hits than bases.

BLAST

BLAST stands for "Basic Local Alignment Search Tool".

It also seeks to exploit the speed of exact pattern matching while factoring in the impact of the scoring matrix.

BLAST also breaks the query into k-mers in order to search a hash table, but for input k-mer q constructs *all* other k-mers which lie within a distance t of q.

By processing all these patterns into an automata, BLAST can then make one linear-time pass through the database to find all exact matches and group them in an alignment.

The window size k is typically 3-5 for protein sequences and 12 for DNA sequences.

Conventional wisdom has BLAST as faster than FASTA, but perhaps a little less accurate.

Using BLAST

Several variants of the *Basic Local Alignment Search Tool* are available at www.ncbi.nlm.nih.gov/BLAST/

blastp	amino acid query sequence to protein sequence database
blastn	nucleotide query sequence to nucleotide sequence database
blastx	nucleotide query sequence translated in all reading frames against a protein sequence database
tblastn	protein query sequence against nucleotide sequence database translated in all reading frames
tblastx	most intensive computationally; compares 6 frame translations of nucleotides query sequence against 6 frame translation of nucleotide sequence database

You can download your own local copy of the code and databases, or use web resources.

BLAST: new tools

PSI-BLAST: Position Specific Iterated BLAST is an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching. A profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. This method is used to increase sensitivity.

PHI-BLAST:Pattern-Hit Initiated BLAST is a search program that combines matching of regular expressions with local alignments surrounding the match. This tool basically searches for motifs.

BLAT: BLAST-Like Alignment Tool is a very fast sequence alignment tool similar to BLAST, which is becoming popular. BLAT is more accurate and can be also much faster than BLAST. BLAT's speed comes from its runtime indexing of all non-overlapping subsequences of given lengths. This index is small enough to fit into computer memory and is typically computed only once for each genome assembly.

BLAST databases

Database choices include:

nr all non-redundant sequences (from particular databases)

est expressed-sequence tags (RNA from expressed genes) in human, mouse, etc.

month new releases from the past 30 days

genomes from Drosophila, yeast, E.coli, human, etc.

Your query sequence can be (1) an amino acid or nucleotide sequence you type or paste in, or (2) the accession or GI number of Genbank entry.

There are a wide range of output formats.

BLAST advanced search parameters

You can select the organism you are interested in to limit your search.

You can change the *expect value* E, the threshold for reporting matches against a database sequence. The default threshold of 10 means that 10 matches are expected to be found merely by chance (Karlin and Altschul). Lower expect thresholds are more stringent, leading to fewer chance matches being reported.

The expect value decreases exponentially with the alignment score S.

You can use *filter* to masks sequences of low compositional complexity, i.e. eliminate statistically significant but biologically uninteresting reports.

You can select the cost comparison matrix. The default is BLOSUM62, but with this matrix fairly long alignments are requires to rise above background. PAM matrices are recommends if search for short alignments.

Significance score

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone.

A model for expected number of *high-scoring segment pairs* (HSPs) with score at least S is

$$E = Kmne^{-\lambda S}$$

where m and n are the sequence lengths and K and λ are scaling parameters.

Doubling length of either sequence doubles E, as it should.

Doubling the score for an HSP to $2x$ requires it to attain the score x twice in a row, so E should decrease exponentially with score.

Significance score

The number of random HSPs with score $\geq S$ is described by a Poisson distribution, so the probability of finding exactly a HSPs with score $\geq S$ is given by $e^{-E}E^a/a!$.

Hence the probability of finding 0 HSPs is e^{-E}, and the probability of finding at least one is $p = 1 - e^{-E}$. This is the *p-value* of the score.

A statistical method to measure significance is to generate many random sequence pairs of the appropriate length and composition, and calculate the optimal alignment score for each.