Experiment 5.2: Homology searches using BLAST and its variants

AIM

To understand and perform sequence similarity searches using BLAST (Basic Local Alignment Search Tool) for both nucleotide sequences (blastn) and protein sequences (blastp), and to learn how different parameters including E-value thresholds, word size, and database selection affect search sensitivity and specificity.

THEORY

Introduction to BLAST

BLAST (Basic Local Alignment Search Tool) is the most widely used bioinformatics program for searching sequence databases to find homologous sequences. Developed by Altschul et al. in 1990, BLAST uses a heuristic algorithm that trades exhaustive searching for speed, making it possible to search massive databases in seconds rather than hours. The fundamental principle behind BLAST is that homologous sequences share discrete conserved regions, and by first identifying these short exact matches (words or seeds), the algorithm can efficiently locate and extend regions of similarity without examining every possible alignment. This approach makes BLAST approximately 50 times faster than dynamic programming methods while maintaining reasonable sensitivity.

BLAST Algorithm and Statistical Significance

The BLAST algorithm operates in three main phases: seeding, extension, and evaluation. In the seeding phase, BLAST breaks the query sequence into short overlapping words (11 nucleotides for blastn, 3 amino acids for blastp by default) and rapidly scans the database for exact or nearly exact matches to these words. The extension phase takes each seed match and extends it in both directions as long as the cumulative alignment score increases or remains above a threshold, creating High-scoring Segment Pairs (HSPs). Finally, the evaluation phase calculates statistical significance for each HSP using the Expect value (E-value), which represents the number of alignments with similar or better scores expected to occur by chance in a database of that size.

The E-value is critical for interpreting BLAST results and depends on both the alignment score and the size of the search space (database size × query length). An E-value of 0.01 means that 0.01 alignments with this score or better would be expected by random chance, suggesting the match is likely biologically significant. Generally, E-values below 1e-5 indicate significant homology, values between 1e-5 and 0.01 suggest possible homology requiring further investigation, and values above 0.1 are likely random matches. The bit score provides a normalized measure independent of database size, making it useful for comparing searches across different databases or time points.

Types of BLAST Programs

NCBI provides several BLAST programs optimized for different types of searches. BLASTn (nucleotide-nucleotide) searches nucleotide databases using nucleotide queries and is ideal for finding identical or highly similar DNA sequences, identifying genes in genomic sequences, or checking for contamination in sequencing projects. BLASTp (protein-protein) searches protein databases using protein queries and is more sensitive than nucleotide searches for detecting distant evolutionary relationships due to the redundancy of the genetic code and the use of sophisticated substitution matrices. BLASTx translates nucleotide queries in all six reading frames to search protein databases, useful for finding protein-coding regions in genomic DNA or EST sequences. tBLASTn searches translated nucleotide databases with protein queries, valuable for finding genes in unannotated genomes, while tBLASTx performs six-frame translations of both query and database, though it is computationally intensive.

Critical Parameters in BLAST Searches

Word size significantly affects search sensitivity and speed, with smaller word sizes increasing sensitivity but requiring more computation time. For blastn, the default word size of 11 works well for similar sequences, while reducing it to 7 increases sensitivity for divergent sequences. For blastp, word sizes of 2 or 3 are typically used, with 2 being more sensitive but slower. The choice of substitution matrix in protein searches affects the ability to detect homologs at different evolutionary distances, with BLOSUM62 serving as the default for standard searches, BLOSUM45 or PAM250 for detecting distant homologs, and BLOSUM80 or PAM30 for closely related sequences.

Database selection is crucial for obtaining relevant results while avoiding spurious matches. Searching the non-redundant (nr/nt) databases provides comprehensive coverage but may return many similar sequences, while organism-specific databases reduce noise and improve relevance for targeted searches. The RefSeq database contains curated sequences with high-quality annotation, making it ideal for functional studies. For preliminary analysis or contamination checking, smaller databases like the 16S ribosomal RNA database or vector databases provide focused results.

EXERCIS: PART A: Identifying an Unknown Gene Using BLASTn

Scenario: You isolated a gene fragment from a bacterial culture and need to identify it.

Steps:

1. Go to https://blast.ncbi.nlm.nih.gov/
2. Click **Nucleotide BLAST**
3. Copy and paste this sequence:

```
>Unknown_Gene_Fragment
ATGGGTAAGGAGGACAAGACTCACCTTAACGTCGTCGTCATCGGCCACGTCGACTCT
GGCAAGTCGACCACTGTAAGTACAACCAACAGCGGGTTGCTTATCTGCACTCGGAAT
CCGCCAAACCTGGCAGGGTATCACCAAAACATCTTGCTAACTTTTGACAGACCGGTCA
CTTGATCTACCAGTGCGGTGGTATCGACAAGCGAACCATCGAGAAGTTCGAGAAGGT
TAGTCAATATCCCTTCGATTACGCGCGCTCCCATCGATTCCCACGATTCGCTCCCTCAC
TCGAAACACATCCATTACCCCGCTCGAGTCCGAAAATTTTGCGGTGCGACCGTGATTT
TTTCTGGTGGGGTATCTTACCCCGCCACTCGAGTCACGGATGCGCTTGCCCTGTTCCC
ACAAAACCTTACCACCCTGTCGCGCACTACATGTCTTGCAGTCACTAACCACTGGACA
ATAGGAAGCCGCCGAGCTCGGAAAGGGTTCCTTCAAGTACGCCTGGGTTCTTGACAA
GCTCAAAGCCGAGCGTGAGCGTGGTATCACCATTGATATCGCTCTCTGGAAGTTCGA
GACTCCTCGCTACTATGTCACCGTCATTGGTATGTTGTCACCGTCTCACACTATCATGT
ATTCATCATGCTAACATCTCTCTCAGATGCCCCCGGTCATCGTGATTTCATCAAGAACA
TGATC
```

4. Click **BLAST**

5. Record Your Findings for the best hit:

- What organism is this gene from?
- What is the gene/protein name?
- What is the E-value of the top hit?
- What percentage identity does it show?
- Is this a significant match (Yes/No)?

EXERCIS: PART B: Finding Human Genes Similar to a Mouse Protein Using tBLASTn

Scenario: You have a mouse protein and want to find similar human genes (including unannotated ones).

Steps:

1. Go to https://blast.ncbi.nlm.nih.gov/
2. Click **tBLASTn** (protein query vs translated nucleotide database)
3. Copy and paste this mouse protein sequence:

   ```
   >Mouse_Myoglobin
   MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSE
   DEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEF
   ISECIIQVLQSKHPGDFGADAQGAMNKALEL FRKDMASNYKELGFQG
   ```

4. In the **Organism** field, type: Homo sapiens (taxid:9606)
5. Click **BLAST**
6. **Record Your Findings:**
    - What is the top human hit?
    - What chromosome is it located on?
    - What is the E-value?
    - What percentage identity to mouse?
    - How many exons are shown in the alignment?

EXERCIS: PART C: Compare Different BLAST Programs:

1. Now go back and try the same search with **BLASTp** instead
2. For BLASTp, change database to "nr" and keep organism as Homo sapiens

Comparison	**tBLASTn Result**	**BLASTp Result**
Type of database searched	Genomic DNA	Protein
Can find unannotated genes?	Yes/No	Yes/No
Number of human hits found		