BLAST: 
Basic Local Alignment Search Tool 

Urmila Kulkarni-Kale 
Bioinformatics Centre 
University of Pune
Sequence based searching

– To compare a sequence against the sequence database

– To locate similar sequences
  • Similarity may extent to entire length
  • Similarity may be restricted to local regions (domains)
Steps in sequence-based database searching

- Identify the query sequence
  - Protein/nucleic acid
- Select an algorithm/tool
  - FASTA / BLAST
- Select the database
  - Protein or nucleic acid sequence database
  - One or all databases
- Fire the query
  - On-line / Off-line
- Analyse the results
  - Statistically significant vs chance findings
DNA vs. Protein searches

• Comparing DNA sequences:
  – More diverged
  – Significantly more random matches
  – No choice of scoring matrices (Unitary matrix)

• Comparing protein sequences
  – Less diverged than the DNA encoding them.
  – Significantly less random hits
  – A wide choice of sensitive matrices like PAM and BLOSUM
Database Searching Programs

- FASTA
- BLAST
- BLITZ
- Smith & Waterman algorithm

Identify local similarity
BLAST Algorithm

(1) For the query find the list of high scoring words of length \( w \).

Query Sequence of length \( L \)

Maximum of \( L-w+1 \) words (typically \( w = 3 \) for proteins)

For each word from the query sequence find the list of words that will score at least \( T \) when scored using a pairscore matrix (e.g. PAM 250). For typical parameters there are around 50 words per residue of the query.
(2) Compare the word list to the database and identify exact matches.

Database Sequences

Exact matches of words from word list
(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold $S$.

Maximal Segment Pairs (MSPs)
# Protein databases for BLAST

<table>
<thead>
<tr>
<th>Database</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nr</strong></td>
<td>Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.</td>
</tr>
<tr>
<td>refseq</td>
<td>Protein sequences from <a href="https://www.ncbi.nlm.nih.gov/">NCBI Reference Sequence project</a>.</td>
</tr>
<tr>
<td>swissprot</td>
<td>Last major release of the SWISS-PROT protein sequence database (no incremental updates).</td>
</tr>
<tr>
<td>pat</td>
<td>Proteins from the Patent division of GenBank.</td>
</tr>
<tr>
<td>month</td>
<td>All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.</td>
</tr>
<tr>
<td>pdb</td>
<td>Sequences derived from the 3-dimensional structure records from the Protein Data Bank.</td>
</tr>
<tr>
<td>env_nr</td>
<td>Non-redundant CDS translations from env_nt entries.</td>
</tr>
<tr>
<td>Smart v4.0</td>
<td>663 PSSMs from Smart, no longer actively maintained.</td>
</tr>
<tr>
<td>Pfam v11.0</td>
<td>7255 PSSMs from Pfam, not the latest.</td>
</tr>
<tr>
<td>COG v1.00</td>
<td>4873 PSSMs from NCBI COG set.</td>
</tr>
<tr>
<td>KOG v1.00</td>
<td>4825 PSSMs from NCBI KOG set (eukaryotic COG equivalent).</td>
</tr>
<tr>
<td>CDD v2.05</td>
<td>11399 PSSMs from NCBI curated cd set.</td>
</tr>
</tbody>
</table>

1: Default; 2: thru rpsblast pages
# Nucleotide databases for BLAST

<table>
<thead>
<tr>
<th>Database</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr</td>
<td>All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer &quot;non-redundant&quot; due to computational cost.</td>
</tr>
<tr>
<td>refseq_mrna</td>
<td>mRNA sequences from NCBI Reference Sequence Project.</td>
</tr>
<tr>
<td>refseq_genomic</td>
<td>Genomic sequences from NCBI Reference Sequence Project.</td>
</tr>
<tr>
<td>est</td>
<td>Database of GenBank + EMBL + DDBJ sequences from EST division.</td>
</tr>
<tr>
<td>est_human</td>
<td>Human subset of est.                                                                隆</td>
</tr>
<tr>
<td>est_mouse</td>
<td>Mouse subset of est.</td>
</tr>
<tr>
<td>est_others</td>
<td>Subset of est other than human or mouse.</td>
</tr>
<tr>
<td>gss</td>
<td>Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.</td>
</tr>
<tr>
<td>htgs</td>
<td>Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.</td>
</tr>
<tr>
<td>pat</td>
<td>Nucleotides from the Patent division of GenBank.</td>
</tr>
<tr>
<td>pdb</td>
<td>Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are NOT the coding sequences for the corresponding proteins found in the same PDB record.</td>
</tr>
<tr>
<td>..</td>
<td>All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the...</td>
</tr>
</tbody>
</table>
BLAST family of programs

• **Blastp**: compares an amino acid query sequence against a protein sequence database

• **Blastn**: compares a nucleotide query sequence against a nucleotide sequence database

• **Blastx**: compares a nucleotide query sequence translated in all reading frames against a protein sequence database

• **Tblastn**: compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames

• **Tblastx**: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
How to run

Input: sequence in FASTA format, Bare sequence, GenBank/GenPept sequence format, copy & paste OR upload as a file OR
Identifiers: accession, accession.version or gi's

Sequence range: 30-300
Specific to protein blast; domain search
Options for advanced BLAST

- Limit the BLAST search to the result of an Entrez query against the database chosen.
  - Mask off segments of the query sequence that have low compositional complexity.
  - Filtering is only applied to the query sequence and not to database sequences.
  - Carried out using SEG and DUST programs.

- Format input sequence to mask certain regions.
  - mRNAs, human repeats, LINE’s, plus retroviral repeats.

- The statistical significance threshold for reporting matches against sequence regions.
BLAST Statistics: significance of E-value

• Quantification of similarity
  – % identity & Similarity score to rank database sequences

• Statistics
  – **E-value** indicates the number of different alignments with score >= S expected to occur by chance in a database search
  – Lower the E-value higher is the significance of score
  – **P-value** indicates if such an alignment can be expected from a chance alone

Chance: can mean the comparison of
(a) real but non-homologous sequences (True negatives)
(b) real sequences that are shuffled to preserve compositional properties
(c) sequences that are generated randomly
Expect value $E()$

– Number of hits expected to be found by chance with a such score.

– $E()$ does not represent a measure of similarity between two sequences.

– As close to 0 as possible
More about E-value

• The number of hits one can "expect" to see just by chance
• Lower the E-value, or the closer it is to "0" the more "significant" the match
• It decreases exponentially with the Score (S) assigned to a match between two sequences.
• For example: E value of 1 assigned to a hit can be interpreted as in a database of the current size one might expect to see 1 match with a similar score simply by chance.
• Note: Searches with short sequences have relatively high E-value meaning shorter sequences have a high probability of occurring in the database purely by chance.
**Test case:** protein

>gi|3328501|Enoyl-Acyl-Carrier Protein Reductase [Chlamydia trachomatis]

MLKIDLTGKIAFIAGDDNGYGWGIAKMLAEAGATILVGTWPIYKIFSQSLELGKFNASRELNGELL
TFAKIYPMDCSFDTPEDIPQEILENKRYKDLGTYTVSEVVEQVKHFGHIDILVHLANSPEIAKPLLDT
SRKGYLAALSTSSYSFISLLSHFGPIMNAGASTISLTLYLASMRAVPGYGGMNAAKAALESDTKVLWEA
GRRWGVNVNTISAGPLASRAGKAIKFIERMVDDYYQDWAPLPSMEMAEQVGAALAVFLVSLASAITGETLY
VDHGAVNVMGIGPEMPKD

- The output
- The first hit
How plant genes were acquired by human parasites?

- *Acanthamoeba*, a free-living protozoan found in fresh water or soil, but which may occur as a human pathogen.
- Perhaps *Acanthamoeba* was the original host for *Chlamydia*, and served as a vector to transfer its *Chlamydia* parasite to humans.
- 16s RNA analyses shows that it is more related to plants

Thus, *Chlamydia* might have acquired plant genes from *Acanthamoeba*
What have we seen?

• A bacterial protein involved in fatty acid metabolism shows similarity with Plant proteins
• The similarity with plant proteins is more than the proteins from other bacteria or the host – human.
• Could it be a case of horizontal gene transfer?
Searching databases

• When searching a database, we take a query sequence and use an algorithm (program) for the search.
• Every pair compared yields a few scores.
• Larger bit/opt scores usually indicate a higher degree of similarity.
• Smaller the E/P values: higher confidence
• A typical db search will yield a huge number of scores to be analyzed.
db searching

- Normally, each database search yields 2 groups of scores: genuinely related (True) and unrelated sequences (False positives), with some overlap between them.
- A good search method should completely separate between the 2 score groups.
- In practice no search method succeeds in total separation.
Sensitivity vs Specificity

• True Positives
• True Negatives
• False Positive: True negative but selected by program as positives
• False Negative: True positive but missed by program and indicated as negative

• Sensitivity:
  – Ability to detect True positive matches
  – Most sensitive search finds all true positives
  – But will also have a few false positives (as low as possible)

• Specificity:
  – Ability to reject True negative matches
  – But will also reject True positives (false negatives)
Sensitivity (Sn) & Specificity (Sp) Calculation

• Sn = TP/ (TP+FN)

• Sp = TP/ (TP+FP)

• Where
  – TP: True Positives
  – FP: False Positive
  – FN: False Negative
Presenting your results

• Document
  – Name and version no of software and database
  – Reference/URL

• Include statistical results that support an inference
  – % identity, P-value, E-value
More BLAST case studies

• Visit Coffee Break @ NCBI