NCBI Molecular Biology Resources

A Field Guide

March 27, 2003

Sequence Similarity Searching

Basic Local Alignment Search Tool

Why do we need similarity searching?

- Identification and annotation
  - Incomplete or no annotations (GenBank)
  - Incorrectly annotated sequences
- Evolutionary relationships
  - Homologous molecules may have similar functions
Molecular Evolution

Common ancestry allows us to infer similar function

- 3000 Myr
- 1000 Myr
- 540 Myr

MLH1 MutL

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation


How BLAST Works

- Make lookup table (hash table) for query
- Scan database for hits
- Ungapped extensions of hits
- Gapped extensions (no traceback)
- Gapped extensions (traceback)
Scoring Systems

- Position Independent Matrices
- Nucleic Acids – identity matrix
- Proteins
  - PAM Matrices (Percent Accepted Mutation)
    - Implicit model of evolution
    - Higher PAM number all calculated from PAM1
    - PAM100 widely used
  - BLOSUM Matrices (BLOCK Substitution Matrices)
    - Empirically determined from alignment of conserved blocks
    - Each includes information up to a certain level of identity
    - BLOSUM62 widely used
- Position Specific Score Matrices (PSSMs)
- PSI and RPS BLAST

BLOSUM62

Position Specific Substitution Rates

[Diagrams and matrices are shown, but the specific content is not transcribed here.]
Position Specific Score Matrix (PSSM)

Gapped Alignments

Scores

V  D  S  -  C  Y
V  E  T  L  C  F
BLOSUM62  +4  +2  +1  -12  +9  +3  7
PAM30  +7  +2  0  -10  +10  +2  11
Web BLAST

BLAST Databases: Non-redundant protein

nr (non-redundant protein sequences)
- GenBank CDS translations
- NP RefSeqs
- Outside Protein
  - PIR Swiss-Prot PRF
- PDB (sequences from structures)
BLAST Databases: Nucleic Acid

- \textit{nr} (nt)
  - Traditional GenBank Divisions
  - NM_ and XM_ RefSeqs
- \textit{dbest}
  - EST Division
- \textit{htgs}
  - HTG division
- \textit{gss}
  - GSS division
- \textit{chromosome}
  - NC_Ref Seqs

---

Mutated in Colon Cancer

IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILER

VQQHIESKLLGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSS

DKVYAHQMVRTDSREQKLDAFLQPLSKPLSS

---

Protein BLAST Page

---

BLAST Formatting Page

Your request has been successfully submitted and put into the Blast Queue.

Query – Mutated in Colon Cancer (131 letters)

Putative conserved domains have been detected, click on the image below for detailed results.

The hit ID is: ZMAM0107411581140

The results are estimated to be ready in 15 minutes but may be done sooner.
BLAST Output: Graphic

Sort by taxonomy

Distributions of Aln Alignment Count on the Query Sequence

BLAST Output: Descriptions

Sort on: protein names, "description" or "score"

BLAST Output: Descriptions

Sort by: taxonomy, e value

TaxBLAST: Taxonomy Reports

Sort on: taxonomy, e value

Sort by: taxonomy, e value
### BLAST Output: Alignments

**Query:** gi|127552|sp|P23367|MUTL_ECOLI DNA mismatch repair protein MutL

**Score:** 44.3 bits (103), Expect = 5e-05
- **Identities:** 25/59 (42%)
- **Positives:** 33/59 (55%)
- **Gaps:** 8/59 (13%)

**Query:**
```
9   LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 59
```

**Subject:**
```
L  +  P   L LEI P  VDVNVHP KHEV F     +H+   + +L  +QQ +E+ L
```

### BLAST Output: Alignments

**Query:** gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair protein Mlh1

**Score:** 233 bits (593), Expect = 4e-61
- **Identities:** 117/131 (89%)
- **Positives:** 117/131 (89%)

**Query:**
```
1   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 60
```

**Subject:**
```
IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 335
```

### Results from nr

Sequences producing significant alignments:
- **Score:** 233 bits (593), Expect = 4e-61
- **Identities:** 117/131 (89%)
- **Positives:** 117/131 (89%)

**Sequence:**
```
IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL
```

**Subject:**
```
IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL
```

-gi|604369|gb|AAA85687.1| (U17857) hMLH1 gene product [Homo sapiens]
-gi|4557757|ref|NP_000240.1| (NM_000249) mutL homolog 1; mutL (E. coli) homolog 1; human homolog of E. coli MutL homolog 1 (colon cancer, nonpolyposis type 2) [Homo sapiens]
-gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair protein Mlh1
-gi|631299|pir||S43085 DNA mismatch repair protein MLH1 - human
-gi|463989|gb|AAC50285.1|(U07343) hMLH1 [Homo sapiens]
-gi|1079787|gb|AAA82079.1|(U40978) DNA mismatch repair protein homolog [Homo sapiens]
-gi|13905126|gb|AAH06850.1|AAH06850 (BC006850) mutL (E. coli) homolog 1 type 2) [Homo sapiens]
-gi|741682|prf||2007430A DNA mismatch repair protein [Homo sapiens]
### tblastn Results Against ESTs

<table>
<thead>
<tr>
<th>Query: 1</th>
<th>GI: 59721448</th>
<th>emb: AB038556.1</th>
<th>Homo sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>prime.</td>
<td>Length = 578</td>
<td>Score = 42 bits (113), Expect(3) = 1e-33</td>
<td>Identities = 10/26 (38%), Positives = 9/26 (35%)</td>
</tr>
<tr>
<td>Frame = +1</td>
<td>Score = 53 bits (138), Expect(3) = 1e-32</td>
<td>Identities = 11/26 (42%), Positives = 11/26 (42%)</td>
<td></td>
</tr>
<tr>
<td>Frame = +2</td>
<td>Score = 53 bits (138), Expect(3) = 1e-32</td>
<td>Identities = 11/26 (42%), Positives = 11/26 (42%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query: 512</th>
<th>GI: 59721448</th>
<th>emb: AB038556.1</th>
<th>Homo sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>prime.</td>
<td>Length = 878</td>
<td>Score = 167 bits (422), Expect(3) = 1e-42</td>
<td>Identities = 81/82 (98%), Positives = 81/82 (98%)</td>
</tr>
<tr>
<td>Frame = +2</td>
<td>Score = 24.3 bits (51), Expect(3) = 1e-42</td>
<td>Identities = 11/26 (42%), Positives = 11/26 (42%)</td>
<td></td>
</tr>
<tr>
<td>Frame = +1</td>
<td>Score = 24.3 bits (51), Expect(3) = 1e-42</td>
<td>Identities = 11/26 (42%), Positives = 11/26 (42%)</td>
<td></td>
</tr>
</tbody>
</table>

---

### PSI-BLAST

**Confirming relationships of purine nucleotide metabolism proteins**

---

### PSI BLAST

**Formatting Options**
- Can be set after the search
- All web BLAST searches are PSI-BLAST
Genomic BLAST pages

Megablast: NCBI's Genome Annotator

- Long alignments between similar DNA sequences
- Concatenation of query sequences
- Faster than blastn
- Discontiguous Megablast allows cross-species comparisons (trace archive and standalone)

Microbial Genomes BLAST

Currently available in BLAST are sequences from reference completed and validated microbial and microbial genomes. Partial genome sequences have been provisionally provided by the sequencing centers through Entrez. For information about the Entrez sequence database, see the link at the bottom of this page. To obtain partial genome sequences data from Entrez, contact the Genome Entrez helpdesk at help@ncbi.nlm.nih.gov.

Enter your query sequence as a text string of FASTA.

Other type of query and databases of NCBI, choose:

- Query: Protein [on], DNA [off]
- Start query: [on], [off]
- All blast [on], [off]
BLAST Results

Map Viewer: Genomic Context of BLAST Hits

NCBI Genomic Resources

Microbial Genomes
Complex Genomes
M. tuberculosis vs. E.coli COGS

Acyl Co-A Dehydrogenase

In Entrez ... coming soon
Human Genome Sequencing Progress

Dec. 2002

Complex Genomes Jan. 2003

- Chordates
  - Human
  - Mouse
  - Rat
  - Pufferfish
  - Sea squirt (Ciona)
- Arthropods
  - D. melanogaster
  - D. simulans
  - A. gambiae
- Higher plants
  - Arabidopsis
  - Rice
- Fungi
  - Aspergillus terreus

Coming soon ...

- In progress
  - purple sea urchin
  - zebrafish
- NHGRI’s Priority Organisms
  - chicken
  - cow
  - dog
  - chimpanzee
  - honeybee
  - Tetrahymena
  - C. elegans
  - several fungi
- Over 100 bacterial genomes
**What Data is Available?**

- Assembled annotated genomic contigs
  - Genome project data
  - Other primary data
- Reference sequences - mRNA, proteins, transcripts
- Genome Scan gene models
- Mapped variation data
- Integrated maps - RH, genetic, cytogenetic, and sequence
- Clustered and mapped expressed sequences
- Links to outside data sources

---

**Higher Genomes Access**

- Entrez databases
- Map Viewer
- Genomic BLAST
- Standard
- UniGene
- LocusLink

---

**LocusLink**

- PubMed
- UCSC
- el
- GDB
- MGC
- HGMD
- ActView
- Homologene
- UniSTS
- Map Viewer
- UniGene
Service Addresses

- General Help  info@ncbi.nlm.nih.gov
- BLAST        blast-help@ncbi.nlm.nih.gov