Dept. of Computer Science and Engineering
University of Rajshahi
www.ru.ac.bd

Dr. Shamim Ahmad

## What is BLAST?

BLAST : Nucleotide/Protein Sequence Databases

as

Google : Internet

## What is BLAST?

Basic Local Alignment Search Tool

Developed in **1989**
National Center for Biotechnology Information **(NCBI)**

**BLAST (Basic Local Alignment Search Tool)**

An algorithm

- For comparing primary biological sequence information,
  - Amino-acid sequences of proteins
  - Nucleotides of DNA sequences.

A BLAST search enables a researcher

- To compare a query sequence with
  - A library
  - Database of sequences
- **Identify library sequences**
  - That resemble the query sequence above a certain threshold.

- BLAST is faster than any Smith-Waterman implementation for most cases
- it cannot "guarantee the optimal alignments of the query and database sequences"

**Input**

Input sequences (in FASTA or Genbank format) and weight matrix.

**Output**

BLAST output can be delivered in a variety of formats

- HTML
- Plain text
- XML

- Query sequence
- Subject sequence
- Specific scoring matrices

DNA or protein input sequence
**Query sequence**
➔ ➔➔ ➔
Database of sequences
**Subject sequences**

Two alignment types are used
- Global
- Local.
- The global approach compares one whole sequence with other entire sequences.

**Global alignment** looks for comparison over the entire range of the two sequences involved.

```
GCATTACTAATATATTAGTAAATCAGAGTAGTA
          ||||||||| ||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

## Local Alignment

- Compares **segments of sequences**
- Finds cases when one sequence is a part of another sequence, or they only match in parts.
- Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. J Mol Biol. 147(1):195-7 (1981)

## What is BLAST?

- BLAST is not Google
- BLAST is like doing an experiment: to get good, meaningful results, you need to optimize the experimental conditions

## Sample Search

- Human beta globin (HBB)
  - Subunit of hemoglobin
- Acquisition number: NP_000509
- Limit to mouse to more easily show differences between searches

## Interpreting Results

- Score: **Normalized score of alignment** (substitution matrix and gap penalty). Can be compared across searches
- Max score: Score **of single best** aligned sequence
- Total score: Sum of scores of all aligned sequences

## Getting the most out of BLAST

1. What kind of BLAST?
2. Pick an appropriate database
3. Pick the right algorithm
4. Choose parameters

## Specialized Search: blastx

- Search **protein** database using a **translated nucleotide** query
- Use to find homologous proteins to a nucleotide coding region

## Specialized Search: tBLASTn

- Search **translated nucleotide** database using a **protein** query
- Does six-frame translations of the nucleotide database
- This program compares a protein query against the all six reading frames of a nucleotide sequence database.

## Specialized Search: tBLASTx

- Search **translated nucleotide** database using a **translated nucleotide** query
- Both translations use all six frames

**Nucleotide-nucleotide BLAST (blastn)**

- This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.

**Protein-protein BLAST (blastp)**

- This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.

**Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)**

- This program is used to find distant relatives of a protein.
- First, a list of all closely related proteins is created.
- These proteins are combined into a general "profile" sequence, which summarizes significant features present in these sequences.

**Nucleotide 6-frame translation-protein (blastx)**

This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence databas

## Step 2: Choose a Database

- Too large:
  - Takes longer
  - Too many results
  - More random, meaningless matches
- Too small or wrong one:
  - Miss significant matches

## Protein Databases

- Non-redundant protein sequences (nr)
  - Kitchen-sink:
    - Translations of GenBank coding sequences (CDS)
    - Ref Seq Proteins
    - PDB (RCSB Protein Data Bank - 3d-structure)
    - SwissProt
    - Protein Information Resource (PIR)
    - Protein Research Foundation (Japanese DB)
- Reference proteins (refseq_protein)
  - NCBI Reference Sequences: Comprehensive, integrated, non-redundant, well-annotated set of sequences
- Swissprot protein sequences (swissprot)
  - Swiss-Prot: European protein database (no incremental updates)

## Protein Databases

- Patented protein sequences (pat)
  - Patented sequences
- Protein Data Bank proteins (pdb)
  - Sequences from RCSB Protein Data Bank with experimentally determined structures
- Environmental samples (env_nr)
  - Protein sequences from environmental samples (not associated with known organism)

## Nucleotide Databases

- Human genomic + transcript
  - http://www.ncbi.nlm.nih.gov/genome/guide/human/
- Mouse genomic + transcript
  - http://www.ncbi.nlm.nih.gov/genome/guide/mouse/
- Nucleotide collection (nr/nt)
  - "nr" stands for "non-redundant," but it isn't
    - GenBank (NCBI)
    - EMBL (European Nucleotide Sequence Database)
    - DDBJ (DNA Databank of Japan)
    - PDB (RCSB Protein Data Bank - 3d-structure)
  - Kitchen sink but not HTGS0,1,2, EST, GSS, STS, PAT, WGS

## Nucleotide Databases

- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
  - NCBI Reference Sequences: Comprehensive, integrated, non-redundant, well-annotated set of sequences
- NCBI Genomes (chromosome)
  - Complete genomes and chromosomes from Reference Sequences

## Nucleotide Databases

- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
  - http://www.ncbi.nlm.nih.gov/About/primer/est.html
  - http://www.ncbi.nlm.nih.gov/dbEST/index.html
- Genomic survey sequences (gss)
  - Like EST, but genomic rather than cDNA (mRNA)
    - random "single pass read" genome survey sequences.
    - cosmid/BAC/YAC end sequences
    - exon trapped genomic sequences
    - Alu PCR sequences
    - transposon-tagged sequences
  - http://www.ncbi.nlm.nih.gov/dbGSS/index.html

## Nucleotide Databases

- High throughput genomic sequences (HTGS)
  - Unfinished sequences (phase 1-2). Finished are already in nr/nt
  - http://www.ncbi.nlm.nih.gov/HTGS/
- Patent sequences (pat)
  - Patented genes
- Protein Data Bank (pdb)
  - Sequences from RCSB Protein Data Bank with experimentally determined structures
  - http://www.rcsb.org/pdb/home/home.do

## Nucleotide Databases

- Human ALU repeat elements (alu_repeats)
  - Database of repetitive elements
- Sequence tagged sites (dbsts)
  - Short sequences with known locations from GenBank, EMBL, DDBJ
- Whole-genome shotgun reads (wgs)
  - http://www.ncbi.nlm.nih.gov/Genbank/wgs.html

## Nucleotide Databases

- Environmental samples (env_nt)
  - Nucleotide sequences from **environmental samples** (not associated with known organism)

## Database Options

- Limit to (or exclude) an organism
- Exclude Models (XM/XP)
  - Model reference sequences produced by NCBI's Genome Annotation project. These records represent the transcripts and proteins that are annotated on the NCBI Contigs … which may have been generated from incomplete data.
- Entrez Query
  - Use Entrez query syntax to limit search

## Step 3: Choose an Algorithm

- How close a match are you looking for?
- Determines how similarities are "scored"
- Affects speed of search and chance of missing match
- Again, what is the goal of the search?

## blastp

- Protein-protein BLAST
- Standard protein BLAST

## megablast

- Nucleotide BLAST
- Finds highly similar sequences
- Very fast
- Use to **identify** a nucleotide sequence

**Highly similar sequences( Mega-BLAST)**

- An algorithm for aligning nucleotide sequences that differ slightly as a result of sequencing or other similar "errors".
- When a larger word size is used, it is up **to 10 times faste**r than more common sequence-similarity programs.
- Mega BLAST can efficiently handle much longer DNA sequences than the traditional BLAST, using the GREEDY algorithm for a nucleotide sequence alignment search.

## blastn

- Nucleotide BLAST
- Use to find less similar sequences

## discontiguous megablast

- Nucleotide BLAST
  - Bioinformatics. 2002 Mar;18(3):440-5. *PatternHunter: faster and more sensitive homology search*. Ma B, Tromp J, Li M.
- Even more dissimilar sequences
- Use to find diverged sequences (possible homologies) from different organisms

NYLENFVQATFN

NYL YLE LEN ENF NFV FVQ VQA QAT ATF TFN

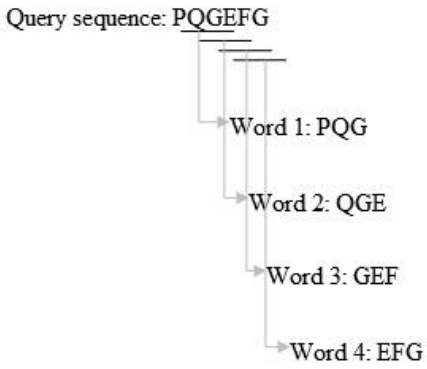
These words are then compared against a sequence in a database.

**Query** ENF
**Subject** SSTNYAENTIQSIISTVEPAQR

The BLAST is a set of algorithms that attempt to find **a short fragment** of a **query sequence** that aligns perfectly with a **fragment of a subject sequenc**e found in a database

For the original BLAST algorithm, the **fragment** is then used as a **seed** to extend the alignment in both directions.

- The first step of the BLAST algorithm is to break the query into short words of a specific length.
- A word is a series of characters from the **query sequences**.
- The default length of the search is **three characters.**
- The words are constructed by using a **sliding window** of three characters.




A small seed is uncovered that can be used to quickly extend the alignment

```
          TAT
          |||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

the extended alignment:

```
          TATATATTAGTA
          ||||||||| ||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

## Alignment

```
AACGTTTCCAGTCCAAATAGCTAGGC
===--===   =-===-==-======
AACCGTTC   TACAATTACCTAGGC
```

Hits(+1): 18
Misses (-2): 5
Gaps (existence -2, extension -1): 1 Length: 3

The hit table includes several useful pieces of information, including the

- **Similarity score**
- **Query coverage**
  - percent of the query sequence that overlaps the subject sequence
- **E-value**
- **Max identity**
  - percent similarity between the query and subject sequences over the length of the coverage area

**BRCA genes**
- Tumor suppressor genes

**BRCA 1**
- Cytogenetic location 17q21
- The q arm of Chromosome 17 at position 21.

**BRCA 2**
- Cytogenetic location 13q12.3
- The q arm of Chromosome 13 at position 12.3.

- Produce proteins that help repair damaged DNA
- Keeping the genetic material of the cell stable. A damaged BRCA gene
- In either location can lead to increased risk of cancer
- Particularly breast or ovarian in women

**BRCA1 and BRCA2 mutations lead**
- preferentially to cancers is not known
- but lack of BRCA1 function seems to lead to non-functional X-chromosome inactivation. Not all mutations are high-risk
- Some appear to be harmless variations

Mutations can be inherited from either parent
- May be passed on to both sons and daughters. Each child of a genetic carrier
- Regardless of sex
- Has a 50% chance of inheriting the mutated gene from the parent who carries the mutation