What is BLAST?

- BLAST stands for
  Basic Local Alignment Search Tool
- Why is BLAST popular?
  - Good balance of sensitivity and speed
  - Reliability
  - Flexibility

Where Can I run BLAST?

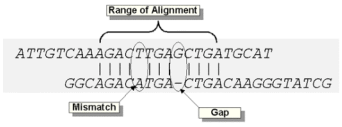
- We will use two sites:
  – NCBI (www.ncbi.nih.gov)
  databases updated constantly (daily); very slow at times
  – FlyBase (http://flybase.org/blast/)
  databases for many *Drosophila* species

BLAST output

- Graphical overview, showing alignment blocks as bars = regions of sequence similarity between Query (top) and database sequences
- List of Sequences with scores (see next slide)
  – Raw score, higher is better (length dependent)
  – Expect value, smaller is better
  (length and database size independent)
- List of alignments

Calculating alignment scores

The raw score S for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. In this figure, a DNA alignment is shown.



$S = \Sigma_{(identities,\ mismatches)} - \Sigma_{(gap\ penalties)}$

$Score = Max(S)$

E value (Expectation value). The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

The Databases

- Genbank nr/nt (protein and nucleotide versions)
  Non-redundant large databases (compile & remove duplicates)
  Anyone can submit, you can call your sequence anything
  Quality low; names can be meaningless
- EST (Expressed Sequence Tags) databases
  Short single reads of cDNA clones
  Other short single reads
  High error rates
- Swissprot
  Curated from literature
  REAL proteins; REAL functions; small;
- Genomic Databases
  Human, Mouse, *Drosophila*, *Arabidopsis*, etc
  NCBI, species-specific web pages

BLAST Protocols

- The most common BLAST search includes five protocols:

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. → Protein
TBLASTN	Nt. → Protein	Protein
TBLASTX	Nt. → Protein	Nt. → Protein

BLASTN

- BLASTN
  - The query is a nucleotide sequence.
  - The database is a nucleotide database
  - No conversion is done on the query or database
- DNA :: DNA homology
  - Mapping oligos to a genome
  - Cross-species sequence exploration
  - Annotating genomic DNA with ESTs

BLASTP

- BLASTP
  - The query is an amino acid sequence
  - The database is an amino acid database
  - No conversion is done on the query or database
- Protein :: Protein homology
  - Protein function exploration
  - Novel gene → makes parameters more sensitive

BLASTX

- BLASTX
  - The query is a nucleotide sequence
  - The database is an amino acid database
  - All six reading frames are translated on the query and used to search the database
- Coding nucleotide seq :: Protein homology
  - Gene finding in genomic DNA
  - Annotating ESTs (and Shotgun Sequence)

TBLASTN

- TBLASTN
  - The query is an amino acid sequence
  - The database is a nucleotide database
  - All six frames are translated in the database and searched with the protein sequence
- Protein :: Coding Nucleotide DB homology
  - Mapping a protein to a genome
  - Mining ESTs (Shotgun DNA) for protein similarities

TBLASTX

- TBLASTX
  - The query is a nucleotide sequence
  - The database is a nucleotide database
  - All six frames are translated on the query and on the database
- Coding :: Coding homology
  - For searching distantly-related species
  - Sensitive but expensive