Searching for similar proteins in a Database

	BLAST	HMM	Threading
Sensitivity:	Least sensitive	→	Most sensitive
Speed:	Seconds	→ Minutes →	Hours
DB size:	1×10^6	→ 1×10^6 →	18000 (PDB)

Searching for similar proteins in a Database

BLAST

http://www.ncbi.nlm.nih.gov/BLAST/

Run blast on your desktop computer

1. Multiple query sequences

2. Specific database that is not available on the NCBI site

3. Customized output format

Project 1: Annotate sequences on a *Drosophila* cDNA microarray

Program: NCBI blast package

Database: *Drosophila* refseq cDNA sequences

Query: EST sequences of the cDNA library

Getting started:

Setting up blast programs on your computer

FTP from: ftp://ftp.ncbi.nih.gov/blast/executables/

Step 1: Setting up blast database

Using NCBI ENTREZ to make a fasta file

URL:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

Search String:

Drosophila[ORGN] AND NM_000000:NM_999999[ACCN]

Save as FASTA file:

dros_refseq

Continue Step 1:

Run formatdb

blast\formatdb -i dros_refseq -p F -o T

(for detailed instructions, see blast\readme.formatdb)

Step 2: Prepare query sequences

Use Batch ENTREZ to make a FASTA file from a list of Accession numbers

Batch ENTREZ URL:
http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide

Accession Number File:
\cbsu\module1\blast_projects\dros_est.acc

Make a fasta file:
dros_est.fa

Step 3: Run blastall:

blast\blastall -p blastn -i dros_est.fa -d dros_refseq -o result1.txt -e 1e-5

blast\blastall -p blastx -i dros_est.fa -d dros_refseq -o result2.txt -e 1e-5 –m 8

Step 4: Parse blast result

parse_blast.pl result1.txt result.xls

Other programs:

1. **blastclust**
   blast\blastclust -i seqfile -S 98

2. **bl2seq**
   blast\bl2seq –i file1 –j file2 –p blastn –o result.txt