Using NCBI BLAST

National Center for Biotechnology Information
Basic Local Alignment Search Tool

NCBI

16 July 2002

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

NCBI

but it ain't necessarily so!

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
- www, email server, standalone, and network clients

NCBI

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)

NCBI

Look Up Table (Hash Table)

Query: **GTQITVEDLFYNIATRRKALKN**

GTQ
TQI
QIT
ITV -> **LTV,MTV,ISV,LSV,MSV**
TVE **IAV,LAV,MAV,ITL,etc.**
VED
EDL
DLF
LFY
FYN

Word Size = 3

Adjustable
2 or 3 for protein (3 default)
> 7 for blastn searches (11 default)

Neighborhood Words

Make table
for both query
and database

NCBI

Messy Details

ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT exact word match

one hit

- Nucleotide BLAST looks for exact matches
- Protein BLAST requires two hits

GTQITVEDLFYNI

SEI **YYN** neighborhood words

two hits

NCBI

More Details (BLAST options)

```
-W  Word size
-f  Threshold for extending hits
-X  X dropoff value for gapped alignment (in bits)
-y  Dropoff (X) for blast extensions in bits
-Z  X dropoff value for final gapped alignment (in bits)
-A  Multiple Hits window size (zero for single hit algorithm)
-e  Expectation value (E)    default = 10.0
-q  Penalty for a nucleotide mismatch (blastn only) default = -3
-r  Reward for a nucleotide match (blastn only)    default = 1
-v  Number of database one-line descriptions
-b  Number of database alignments
```

NCBI

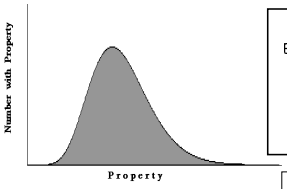
An alignment that BLAST can't find

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

NCBI

Local Alignment Statistics

High scores of local alignments between two random sequences follow Extreme Value Distribution



For ungapped alignments:

Expected number with score S or greater

$E = Kmne^{-\lambda S}$

or

$E = mn2^{-S'}$

K = scale for search space
λ = scale for scoring system
S' = bitscore = $(\lambda S - \ln K)/\ln 2$

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

NCBI

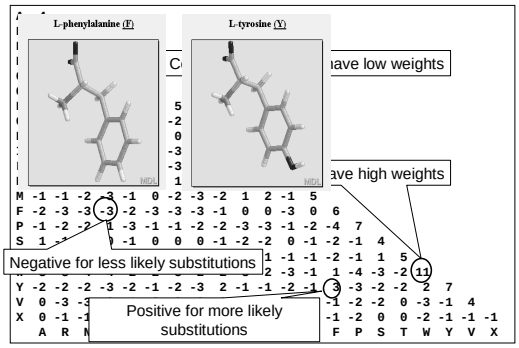
Scoring Systems

	A	G	C	T
A	+1	−3	−3	−3
G	−3	+1	−3	−3
C	−3	−3	+1	−3
T	−3	−3	−3	+1

- Nucleic acids identity matrix
- Proteins
  - Position Independent Matrices
    - PAM Matrices (Percent Accepted Mutation)
      - Implicit model of evolution
      - Higher PAM number all calculated from PAM1
      - PAM250 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

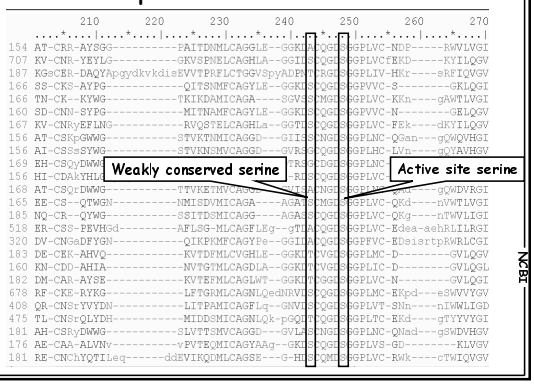
NCBI

BLOSUM62



NCBI

Position Specific Substitution Rates



NCBI

Position Specific Score Matrix (PSSM)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206 D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207 G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208 V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209 I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210 S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211 S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212 C	-4	-7	-6	-7	12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-7	-4	-4	-5	0	-4
213 N	-2	0	2	-1	-6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	-1	-3	-3	-4	-3
214 G	-2	-3	-3	-4	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-6	-3	-5	-6	-6	-6
215 D	-5	-5	-2	9	-7	-4	-1	-5	-5	-7	-7	-4	-7	-7	-5	-4	-4	-8	-7	-7
216 S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217 G	-3	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218 G	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-7	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219 P	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-6	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220 L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221 N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222 C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223 Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224 A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

NCBI

Gapped Alignments

- Gapping provides more biologically realistic alignments
- Statistical behavior not completely understood for gapped alignments
  - Gapped BLAST parameters must be found by simulations for each matrix
- Affine gap costs = -(a+bk)
  - a = gap open penalty b = gap extend penalty
  - A gap of length 1 receives the score -(a+b)

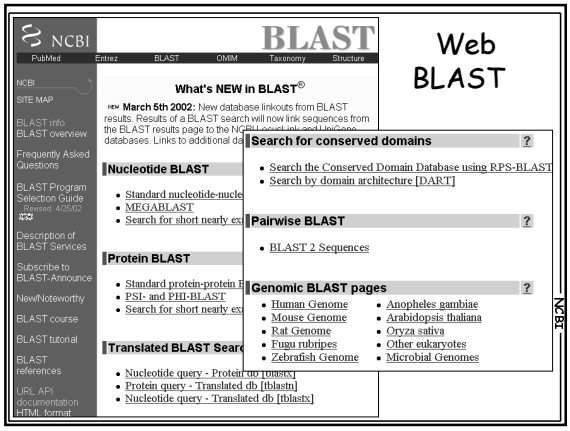
NCBI

Scores

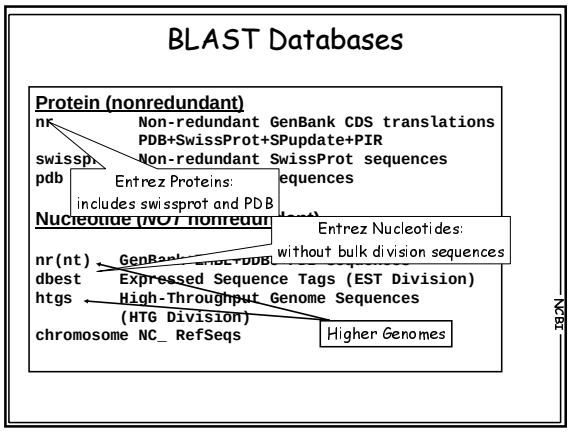
	V	D	S	–	C	Y	
	V	E	T	L	C	F	
BLOSUM62	+4	+2	+1	-12	+9	+3	7
PAM30	+7	+2	0	-10	+10	+2	11

NCBI

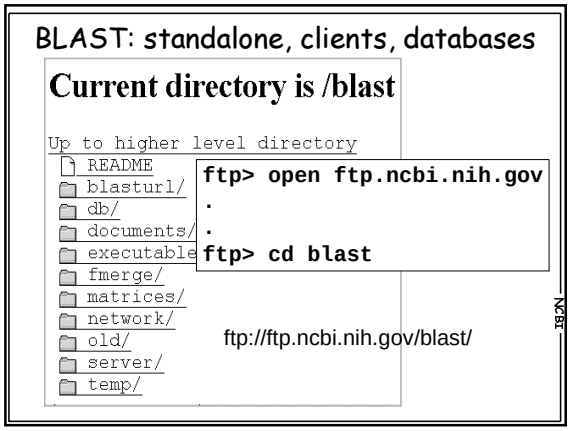
Web BLAST



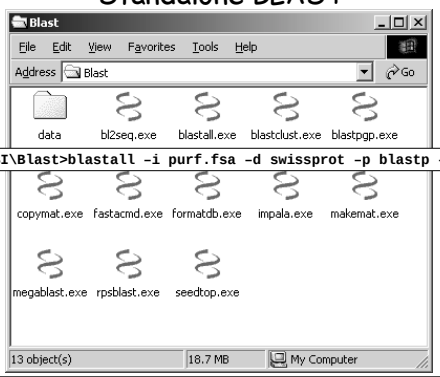
BLAST Databases



BLAST: standalone, clients, databases



Standalone BLAST




```
C:\NCBI\Blast>blastall -i purf.fsa -d swissprot -p blastp -o out
```

NCBI

BLAST Batch Client

```
C:\Netblast>blastcl3 -i input.seq -d nr -p blastn -o outfile
```

```
National Center for Biotechnology Information (NCBI)

welcome to the blast network service.
BLASTN 2.2.2 [Dec-14-2001]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|19343352|gb|AAF15280.2|AF192502_1 aryl hydrocarbon receptor
[Gallus gallus]
         (535 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,
or phase 0, 1 or 2 HTGS sequences)
           1,214,651 sequences; 1,071,392,519 total letters
```

NCBI