# Basic Local Alignment Search Tool

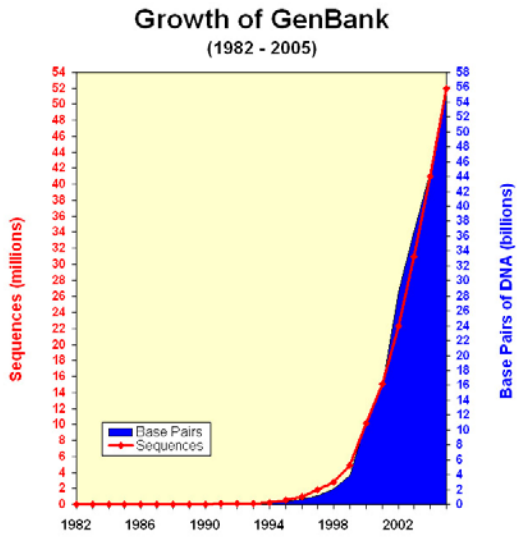
## A blast from the past…

AGATCAC
CGACAG

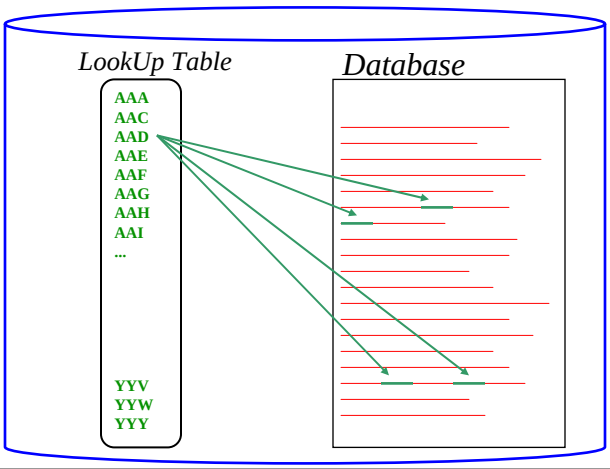
| | | A | G | A | T | C | A | C |
|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 5 |
| G | 0 | 0 | 5 | 0 | 0 | 0 | 1 | 0 |
| A | 0 | 5 | 0 | 10 | 4 | 0 | 5 | 0 |
| C | 0 | 0 | 1 | 4 | 6 | 9 | 3 | 10 |
| A | 0 | 5 | 0 | 6 | 0 | 3 | 14 | 8 |
| G | 0 | 0 | 10 | 4 | 2 | 0 | 8 | 10 |

GATCA
|| ||
GA-CA

## Why BLAST?



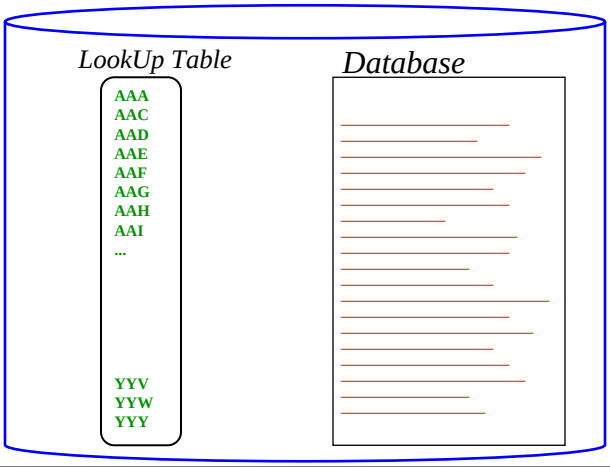

## While you were sleeping...



## BLAST Example

*Query sequence*

**MLVFAHAYHESKWAAHNQEILTPLV**



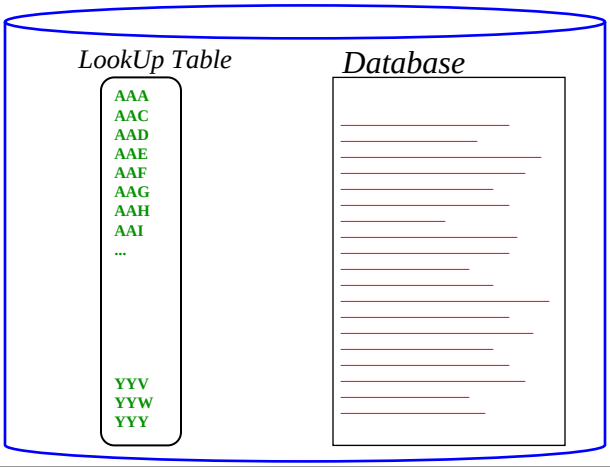

## BLAST Example

*Query sequence*

**MLVFAHAYHESKWAAHNQEILTPLV**

*Word List*

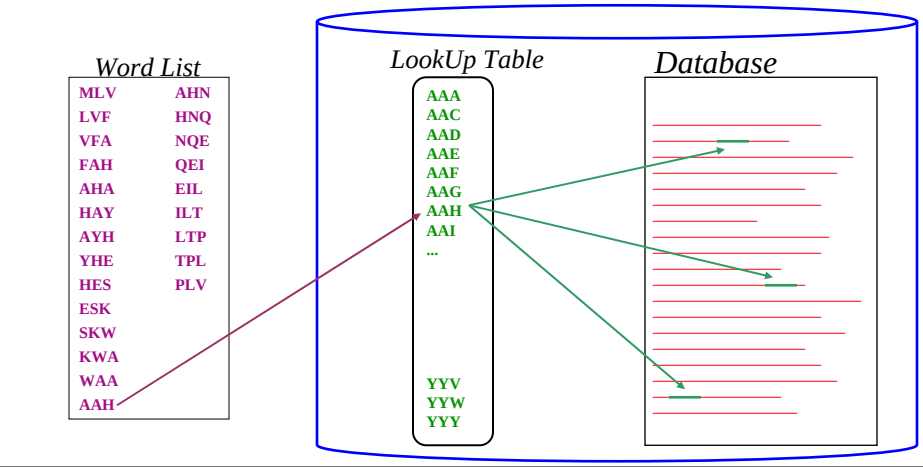
| | |
|---|---|
| MLV | AHN |
| LVF | HNQ |
| VFA | NQE |
| FAH | QEI |
| AHA | EIL |
| HAY | ILT |
| AYH | LTP |
| YHE | TPL |
| HES | PLV |
| ESK | |
| SKW | |
| KWA | |
| WAA | |
| AAH | |



## BLAST Example

*Query sequence*

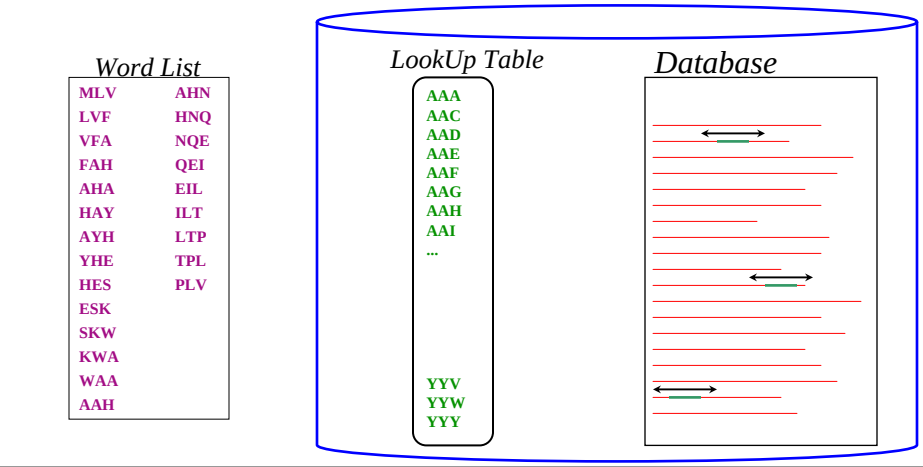
**MLVFAHAYHESKWAAHNQEILTPLV**




## BLAST Example

*Query sequence*

**MLVFAHAYHESKWAAHNQEILTPLV**

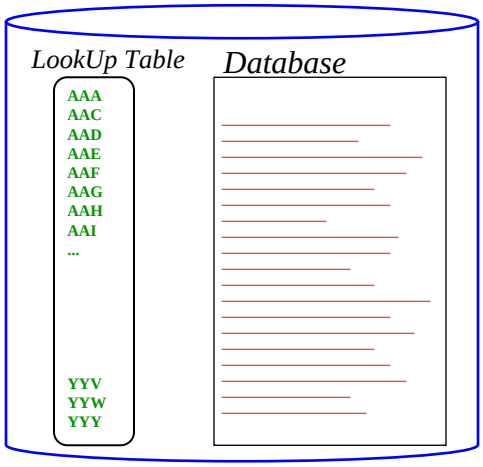


# BLAST In a Nutshell

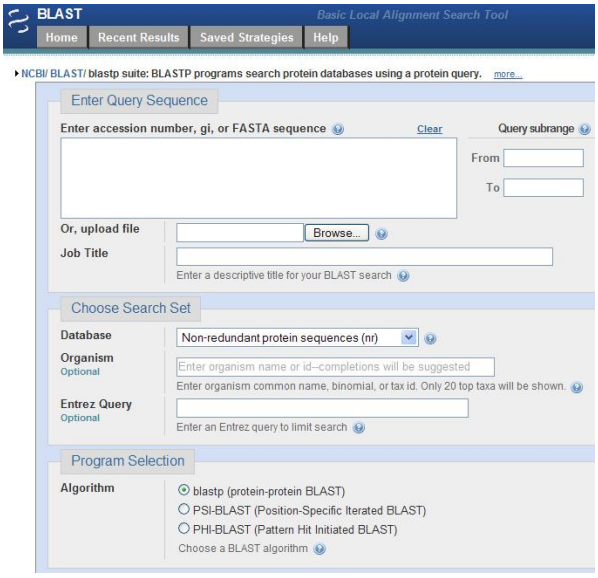
*Query sequence*

**MLVFAHAYHESKWAAHNQEILTPLV**

- Create "word list" from query sequence
- Locate *words* in database via "lookup table"
- Determine similarity of query sequence to each word-match sequence in database




# BLAST Program



# BLAST Output

```
dbj|BAA29916| (AP000003) 170aa long hypothetical protein [Pyrococcus horikoshii]
          Length = 170

 Score =  107 bits (264), Expect = 6e-23
 Identities = 63/160 (39%), Positives = 97/160 (60%), Gaps = 7/160 (4%)

Query: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60
           M  M++K+L+PTDFSE A  A++  +    ++  EVILLHVIDE  +++     L+ G +
Sbjct: 1   MIFMFRKVLFPTDFSEGAYRAVEVFEKRNKMEVGEVILLHVIDEGTLEE-----LMDGYS 55

Query: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDV--GFKVKDIIVVGIPHEEIVKIAED 118
               + E    ++K KL EEA  K++   +E++       V+ II  GIP +EIVK+AE+
Sbjct: 56  FFYDNAEIELKDIKEKLKEEASRKLQEKAEEVKRAFRAKNVRTIIRFGIPWDEIVKVAEE 115

Query: 119 EGVDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVK 158
           E V +II+ S GK +L    LGS    V++K+ KPVL++K
Sbjct: 116 ENVSLIILPSRGKLSLSHEFLGSTVMRVLRKTKKPVLIIK 155
```

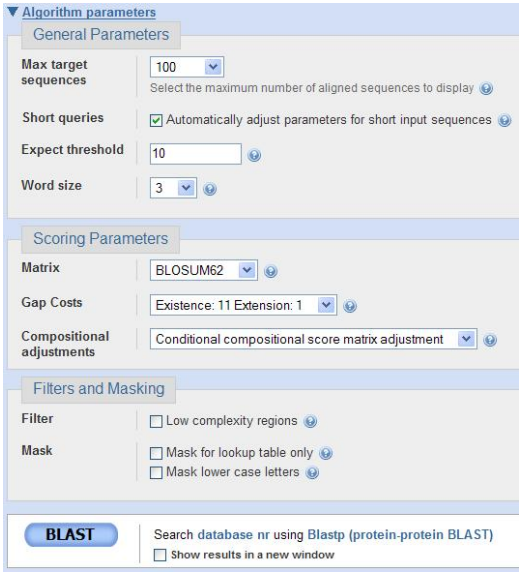
# BLAST Output

**Step 6. Statistical details of the search** Details

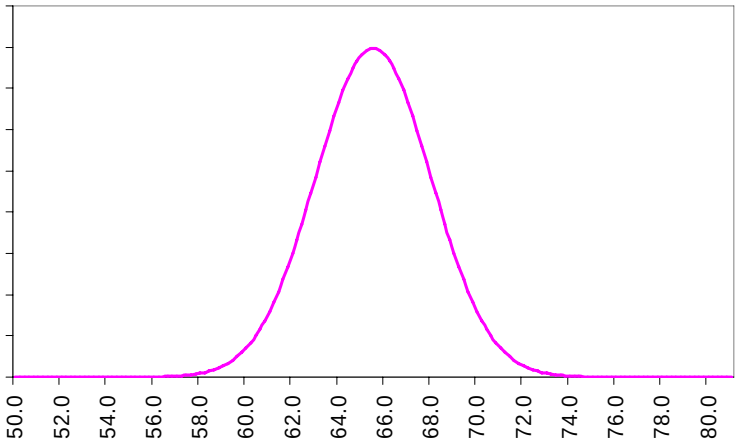
1. Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
2. Posted date: Feb 29, 2008 6:04 PM
3. Number of letters in database: 2,144,987,218
   Number of sequences in database: 6,276,778
4. Lambda K H
   0.314 0.135 0.352

   Gapped Lambda K H
   0.267 0.0410 0.140
5. Matrix: BLOSUM62
6. Gap Penalties: Existence: 11, Extension: 1

## BLAST Options

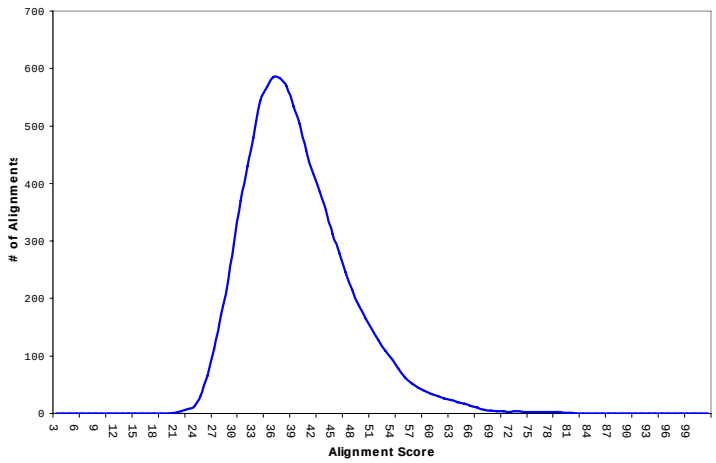



## Normal Distributions




The heights of women are normally distributed, with a mean of 65.5 inches and a standard deviation of 2.5 inches.

## Extreme Value Distributions




Scores of optimal local alignments correspond to extreme value distributions.

## Statistical Significance

Suppose we align two sequences, a query sequence and a target sequence, and we determine that their optimal local alignment score is $S = 60$.

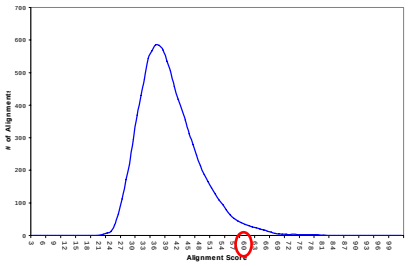
Are the sequences similar? In other words, is a score of $S = 60$ significant? How likely is it that we would observe an alignment score of $S = 60$ by chance?

The *p*-value of an optimal local alignment score, $S$, is the likelihood that two random sequences* would have an optimal local alignment score greater than or equal to $S$.

* of the same lengths and compositions as the query and target sequences

## *p*-values for pairs of sequences

What is the probability that the optimal local alignment score of two sequences will be at least 60?




Solution 1: Count up all of the alignment scores greater than or equal to 60 and divide by the total number of alignment scores, i.e., 10,000.

Solution 2: Plug $x = 60$ into the the following expression, where $\mu = 34.2$ and $\beta = 6.1$

$$1.0 - e^{-e^{-\frac{x-\mu}{\beta}}}$$

## *p*-values for databases

When searching a large database with many target sequences, our previous definition of the *p*-value is problematic because we can expect some small *p*-values by chance. For example, if we align a query sequence to 6,000,000 target sequences in a database, we can expect 60,000 scores with a *p*-value less than 0.01.

When we BLAST a query sequence against a database of many target sequences, the *p*-value of one of the alignment scores, *S*, indicates the likelihood that we would see a score of at least *S* when BLASTing the query sequence against a comparable random database.

## E-values

Instead of *p*-values, BLAST reports E-values. If the alignment score of a query sequence and some target sequence in the database is *S*, the E-value is the expected number of alignments with score *S* or higher in a random database.

```
Query  196   AHAYEGYQDRLILREGLLVALFLAGLVILGGQQQWWLQPVLLGMTSDQVFFGAAILTAFT  255
             A  Y    + L++  G L    L G V+  GQ    L P LL      + FFG  +
Sbjct  1175  AQLYGDTLNLLVIDSGRLAINCLIGFVLFNGQPASQLPPRLLEDAEYKSFFGTRMFPVAV  1234

Query  256   DNAALTYLGSLVAGLSDGFKYALV--AGAVTGGGLTIM----------ANAPNPAGIAIL  303
             D    TY    VA + DGF Y  +   GA  G  L+I+          A  P P    +L
Sbjct  1235  DKDGKTYRA--VAPI-DGFMYTFIDIDGAADGHQLSIIQEPEPKPDDPAGGPRPGQRLLL  1291

Query  304   -RGHFKG-ASVHPL  315
              RG  +   S HPL
Sbjct  1292  ARGLARALLSEHPL  1305
```

## E-values depend on sequences and scoring