Alignment Search Space

```
ABC
XYZ
```

```
A-BC
XYZ-
```

```
A-BC
-XYZ
```

```
AB-C
X-YZ
```

```
AB-C
XYZ-
```

```
AB-C
-XYZ
```

```
---ABC
XYZ---
```

```
ABC-
-XYZ
```

```
ABC--
--XYZ
```

- L_1*L_2
- 200*200 = 40,000 alignments
- Intractable to sample all possible alignments
- *Dynamic programming* reduces the search space

Pairwise Sequence Alignment: Summary

- Dynamic programming
  - Local alignment (Smith-Waterman)
  - Global Alignment (Needleman-Wunsch)
- Substitution matrix: represents the propensity of amino acids to substitute each other over evolutionary time

BLAST

- Used to search large databases with a single sequence
- Basic Local Alignment Search Tool
- 1993 / 1997 Samuel Karlin and Steven Altschul
- Popular. Most cited paper in molecular biology.
- BLAST is an acceptable verb in scientific papers

Why BLAST?

- Faster than ~~a speeding bullet~~ dynamic programming
- Accurate enough
- Solid hypothesis statistics
- Free code: anyone can alter (WU-BLAST, BLAT)

Uses for BLAST

- Find gene/protein family members
- Predict the function of a gene
- Finding a protein family member
- Predicting a protein's 3D structure

How BLAST works

1) List of high scoring words in query
2) Scan the sequence database
3) Extend hits
4) Rank & report

Step 1: High Scoring Words

Step 1: Compile a list of high-scoring words above threshold T.

Query sequence: . . . RCPHHERCSD. . .

Words derived from query sequence: RCP, CPH, PHH, HHE, …

List of neighboring words to RCP above threshold T=16:

Word	Scores from BLOSUM scoring matrix	Total score
RCP	5 + 9 + 7	21
KCP	2 + 9 + 7	18
QCP	1 + 9 + 7	17
ECP	0 + 9 + 7	16
.	.	.
.	.	.

Note: The line is located at the threshold.
Word size is 3.

Steps 2&3

Step 2: Scan the database for short segments that match the list of acceptable words/scores above or equal to threshold T.

Step 3: Extend the hits and terminate when the tabulated score drops below a cutoff score.

```
               ←─────    ──────────→
Query           EVVRRCPHHERCSD
                EVVRRCPHHER S+
Sbjct           EVVRRCPHHERSSE
```

If the hit is extended far enough, the query/subj segment is called a High Scoring Segment Pair (HSP).

Step 4: rank results

```
>sp|P96681.1|YDFD_BACSU  RecName: Full=Uncharacterized HTH-type transcriptional
regulator
ydfD
Length=482
 Score = 34.7 bits (78),  Expect = 1.3, Method: Compositional matrix adjust.
 Identities = 32/150 (21%), Positives = 64/150 (42%), Gaps = 15/150 (10%)
Query  130  GALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDTCALDFGGLI  189
            GAL+          Q    +Y+  PS+   + +F+ AG+ +            +D  GL+
Sbjct  189  GALQALQLISMGLLQRGSTVYLDQPSYLYSLHVFQSAGMKLT--------GLPMDNEGLL  240
Query  190  EDLKKIPE----KSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGFAT  245
             +   +      ++I+  + C HNPTG+  + ++  EI A+ +   L    D  Y+
Sbjct  241  PENVHLTRGERGRAILYTNPCFHNPTGILMSKKRREEILAVSENTQLPIIEDDIYRELWI  300
Query  246  GDIDRDAQAVRTFEADGHDFCLAQSFAKNM  275
             +I      ++T + +GH   +  S +K +
Sbjct  301  DEI--PPYPIKTIDKNGHVLYIG-SLSKTL  327


>sp|Q93ZN9.1|DAPAT_ARATH  RecName: Full=LL-diaminopimelate aminotransferase,
chloroplastic;
Short=LL-DAP-aminotransferase; Short=DAP-aminotransferase;
Short=DAP-AT; Short=AtDAP-AT; AltName: Full=Protein ABERRANT
GROWTH AND DEATH 2; Flags: Precursor
Length=461
 Score = 34.7 bits (78),  Expect = 1.3, Method: Compositional matrix adjust.
 Identities = 28/99 (28%), Positives = 38/99 (38%), Gaps = 8/99 (8%)
Query  187  GLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGFATG  246
            G   DL  +    I+      +NPTG   T EQ  ++    KK       D AY  + +
Sbjct  223  GFFPDLSTVGRTDIIFF-CSPNNPTGAAATREQLTQLVEFAKKNGSIIVYDSAYAMYMSD  281
Query  247  DIDRDAQAVRTFEADGHDFCLAQ--SFAKNMGLYGERAG  283
            D  R       FE  G +    +  SF+K  G  G R G
Sbjct  282  DNPRS-----IFEIPGAEEVAMETASFSKYAGFTGVRLG  315
```

What are the different BLAST programs?

- blastp
  - compares an amino acid query sequence against a protein sequence database
- blastn
  - compares a nucleotide query sequence against a nucleotide sequence database
- blastx
  - compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- tblastn
  - compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- tblastx
  - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

What are the different BLAST programs? (continued)

- psi-blast
  - Compares a protein sequence to a protein database. Performs the comparison in an iterative fashion in order to detect homologs that are evolutionarily distant.
- blast2
  - Compares two protein or two nucleotide sequences.

Let's BLAST stuff

http://www.ncbi.nlm.nih.gov/Class/minicourses/quickblast.html

http://www.ncbi.nlm.nih.gov/books/NBK1734/