UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Bioinformatics Tutorials

Algorithms for Sequence Analysis

Iosif Vaisman

2000

Comparative Sequence Sizes

- Yeast chromosome 3 350,000
- Escherichia coli (bacterium) genome 4,600,000
- Largest yeast chromosome now mapped 5,800,000
- Entire yeast genome 15,000,000
- Smallest human chromosome (Y) 50,000,000
- Largest human chromosome (1) 250,000,000
- Entire human genome 3,000,000,000

The String Alignment Problem

string - a sequence of characters from some alphabet

given: two strings **acbcdb** and **cadbd**

one of possible alignments:

```
a c - - b c d b
- c a d b - d -
```

scoring function:
exact match +2
mismatch -1
insertion -1

score:
$3 \cdot (2) + 5 \cdot (-1) = 1$

The String Alignment Problem

given: two strings **CTCATG** and **TACTTG**

```
C T C A T G
    |   | |
T A C T T G
```

score:
$3 \cdot (2) + 3 \cdot (-1) = 3$

```
C T C A - T - G
  |   |   |   |
· T - A C T T G
```

score:
$4 \cdot (2) + 4 \cdot (-1) = 4$

Entropy and Redundancy of Language

```
  CUR     F     W       D      DIS     AND P
A     SED   IEND  ROUGHT   EATH     EASE      AIN
  BLES    FR    B       BR     AND         AG
```

Entropy and Redundancy of Language

```
  CUR     F     W       D      DIS     AND P
A     SED   IEND  ROUGHT   EATH     EASE      AIN
  BLES    FR    B       BR     AND         AG
```

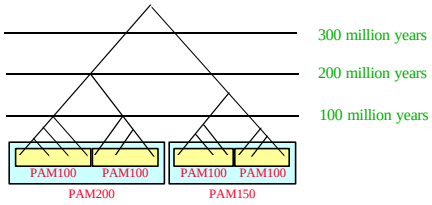
The central raw contains 65% of the letters, but does not affect the content of messages

```
A  CURSED  FIEND WROUGHT  DEATH  DISEASE AND PAIN
||   ||||  ||||| |||||||  |||||   |||||     |||
A BLESSED FRIEND BROUGHT BREATH AND EASE    AGAIN
```

Substitution Matrices

- Dayhoff (or MDM, or PAM) -
  Derived from global alignments of closely related sequences

PAM100 - number referes to evolutionary distance
(Percentage of Acceptable point Mutations per 10^8 years)



Substitution Matrices

- BLOSUM (BLOcks SUbstitution Matrix) -
  Derived from local, ungapped alignments of distantly related sequences

BLOSUM62 - number refers to the minimum percent identity

Reference: Henikoff & Henikoff *Proteins* **17**:49, 1993

Selecting a Matrix

- Compared sequences are related:
  200 PAM or 250 PAM
- Database scanning:
  120 PAM
- Local alignment search:
  40 PAM, 120 PAM, 250 PAM
- Detection of related sequences using BLAST:
  BLOSUM 62

Low PAM:
short segments,
high similarity

High PAM:
long segments,
low similarity

THERE IS NO "ONE SIZE FITS ALL" MATRIX !

Matrix Example

```
 A   B   C   D   E   F   G   H   I   K    ..
1.5 0.2 0.3 0.3 0.3 -0.5 0.7 -0.1 0.0 0.0   .. A
    1.1 -0.4 1.1 0.7 -0.7 0.6 0.4 -0.2 0.4  .. B
        1.5 -0.5 -0.6 -0.1 0.2 -0.1 0.2 -0.6 .. C
            1.5 1.0 -1.0 0.7 0.4 -0.2 0.3   .. D
                1.5 -0.7 0.5 0.4 -0.2 0.3   .. E

                    1.5 -0.6 -0.1 0.7 -0.7  .. F
                        1.5 -0.2 -0.3 -0.1  .. G
                            1.5 -0.3 0.1    .. H
                                1.5 -0.2    .. I
                                    1.5     .. K
```

Dayhoff's Acceptable Point Mutations

```
Ala A
Arg R  30
Asn N 109  17
Asp D 154   0 532
Cys C  33  10   0   0
Gln Q  93 120  50  76   0
Glu E 266   0  94 831   0 422
Gly G 579  10 156 162  10  30 112
His H  21 103 226  43  10 243  23  10
Ile I  66  30  36  13  17   8  35   0   3
Leu L  95  17  37   0   0  75  15  17  40 253
Lys K  57 477 322  85   0 147 104  60  23  43  39
Met M  29  17   0   0   0  20   7   7   0  57 207  90
Phe F  20   7   7   0   0   0   0  17  20  90 167   0  17
Pro P 345  67  27  10  10  93  40  49  50   7  43  43   4   7
Ser S 772 137 432  98 117  47  86 450  26  20  32 168  20  40 269
Thr T 590  20 169  57  10  37  31  50  14 129  52 200  28  10  73 696
Trp W   0  27   3   0   0   0   0   0   3   0  13   0   0  10   0  17   0
Tyr Y  20   3  36   0  30   0  10   0  40  13  23  10   0 260   0  22  23   6
Val V 365  20  13  17  33  27  37  97  30 661 303  17  77  10  50  43 186   0  17
      A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y
    Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr
```

Search and alignment entropy

- **Information content per position:**
  pam10 - 3.43 bits
  pam120 - 0.98 bits
  pam160 - 0.70 bits
  pam250 - 0.38 bits
  blosum62 - 0.70 bits

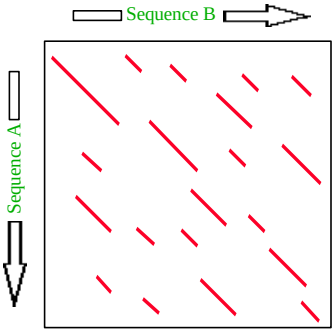
- **Information requirements:**
  for search - 30 bits
  for alignment - 16 bit

Search and alignment entropy

Recommended matrices for different query length

Query length	Substitution matrix	Gap costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1)

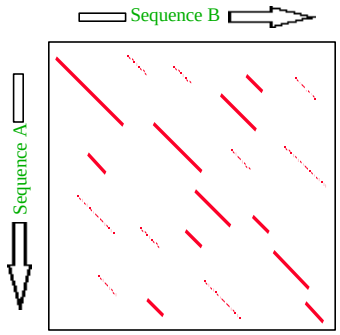
FASTA Algorithm



1

First run (identities)

FASTA Algorithm



2

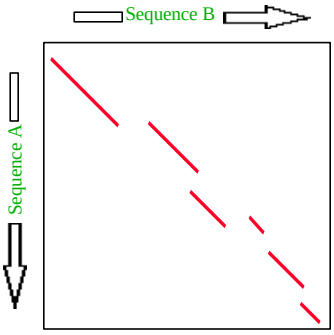
Rescoring using PAM matrix

high score

low score

The score of the highest scoring initial region is saved as the **init1 score**.

FASTA Algorithm

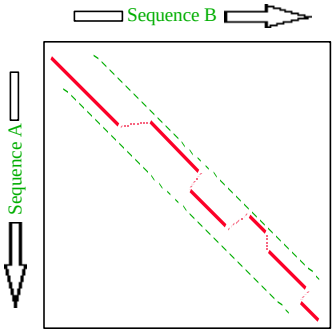


3

Joining threshold - eliminates disjointed segments

Non-overlapping regions are joined. The score equals sum of the scores of the regions minus a gap penalty. The score of the highest scoring region, at the end of this step, is saved as the **initn score**.

FASTA Algorithm



4

Alignment opyimization using dynamic programming

The score for this alignment is the **opt score**.

FASTA Algorithm

FastA uses a simple linear regression against the natural log of the search set sequence length to calculate a normalized **z-score** for the sequence pair.

Using the distribution of the z-score, the program can estimate the number of sequences that would be expected to produce, purely by chance, a z-score greater than or equal to the z-score obtained in the search. This is reported as the **E() score**.

FASTA Results

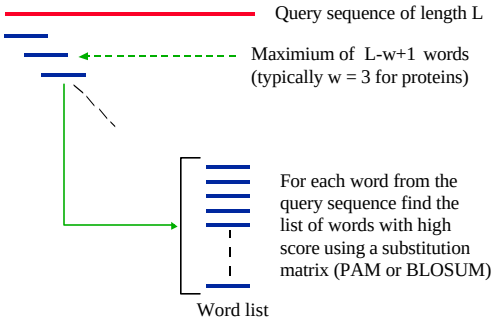
- When init1=init0=opt:
  100 % homology over the matched stretch.
- When initn > init1:
  more than 1 matching region in the database with poorly matching separating regions.
- When opt > initn:
  the matching regions are greatly improved by adding gaps in one or both of the sequences.

BLAST - Basic Local Alignment Search Tool

- Blast programs use a heuristic search algorithm. The programs use the statistical methods of Karlin and Altschul (1990,1993).
- Blast programs were designed for fast database searching, with minimal sacrifice of sensitivity to distant related sequences.

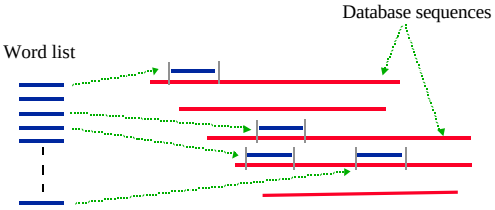
BLAST Algorithm

1



BLAST Algorithm

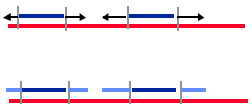
2



Exact matches of words from the word list to the database sequences

BLAST Algorithm

3



Maximal Segment Pairs (MSPs)

For each exact word match, alignment is extended in both directions to find high score segments

Gapped BLAST

- The Gapped Blast algorithm allows gaps to be introduces into the alignments. That means that similar regions are not broken into several segments.
- This method reflects biological relationships much better.

BLAST family of programs

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

Database Searches

- Run Blast first, then depending on your results run a finer tool (Fasta, Smith-Waterman, etc.)
- Where possible use translated sequence.
- E() < 0.05 is statistically significant, usually biologically interesting. Check also 0.05 < E() <10 because you might find interesting hits.
- Pay attention to abnormal composition of the query sequence, it usually causes biased scoring.
- Split large query sequence (if >1000 for DNA, >200 for protein).
- If the query has repeated segments, remove them and repeat the search.

Documenting the Search

- Algorithm(s)
- Substitution matrix
- Gap penalty (FASTA)
- Name of database
- Version of database
- Computer used