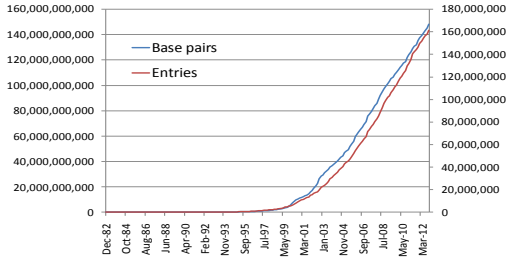
Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

Growth of GenBank



December 1982
680,338 bp
606 seq

December 2012
148,390,863,904 bp
161,140,325 seq

Comparative Sequence Sizes

- Smallest bacterial gene 54
- Smallest human gene (IGF-2) 252
- Yeast chromosome 3 350,000
- Longest human gene (dystrophin) 2,220,223
- 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

two strings **acbcdb** and **cbabd**
can be aligned in different ways

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

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

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

The String Alignment Problem

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

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

score:
$3 \cdot (2) + 4 \cdot (-1) = 2$

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

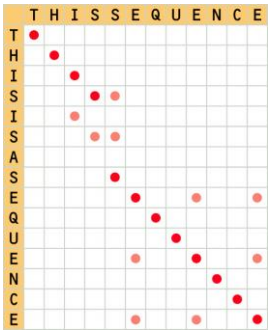
```
** CUR**** F******W******* D***** DIS*****AND P***
||    ||||  |||||| |||||||  |||||    |||||     |||
**BLES****FR******B*******BR******AND *****   AG***
```

The sequences are 65% identical

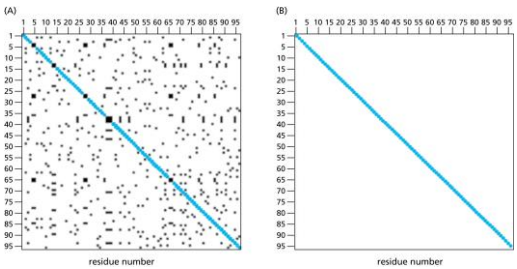
```
A  CURSED  FIEND WROUGHT  DEATH  DISEASE AND PAIN
||    ||||  |||||| |||||||  |||||    |||||     |||
A BLESSED FRIEND BROUGHT BREATH AND EASE    AGAIN
```

Scoring Alignments

```
THISSEQUENCE
||   ||||||||
THATSEQUENCE

THISSEQUENCE
||||||||||||
THISSEQUENCE


THIS-S--EQUENCE
||||  |  |||||||
THISISASEQUENCE

THIS---SEQUENCE
||||    ||||||||
THISISASEQUENCE
```



Scoring Alignments

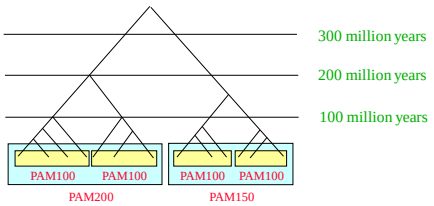



```
WYFGKITRRESERLLLNPENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKL
WYFGKITRRESERLLLNPENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKL
```

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

Specialized Substitution Matrices

- SLIM (ScoreMatrix Leading to Intra-Membrane)
- PHAT (Predicted Hydrophobic and Transmembrane Matrix)
- STROMA (Score Matrix for Known Distant Homologs)
- HSDM (Homologous Structure Derived Matrix)
- WAC (Amino Acid Comparative Profiles)
- VTML (Maximum Likelihood Estimation)

Comprehensive list: Amino Acid Matrices in AAindex
(http://www.genome.jp/aaindex/AAindex/list_of_matrices)

Selecting a Matrix

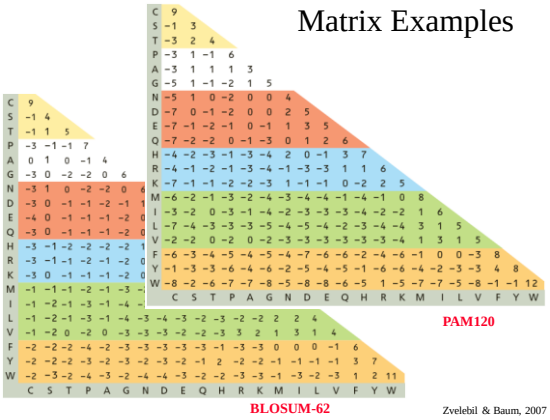
- Compared sequences are related:
  200 PAM or 250 PAM
- Database scanning:
  120 PAM
- Local alignment search:
  40 PAM, 120 PAM, 250 PAM
- Detection of related sequences using BLAST:
  BLOSUM 62

Low PAM:
short segments,
high similarity

High PAM:
long segments,
low similarity

THERE IS NO "ONE SIZE FITS ALL" MATRIX !

Matrix Examples



Zvelebil & Baum, 2007

Search and alignment entropy

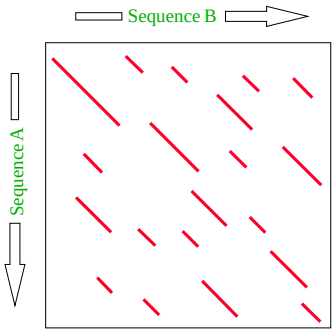
- **Information content per position:**
  - pam10 - 3.43 bits
  - pam120 - 0.98 bits
  - pam160 - 0.70 bits
  - pam250 - 0.38 bits
  - blosum62 - 0.70 bits
- **Information requirements:**
  - for search - 30 bits
  - for alignment - 16 bit

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)

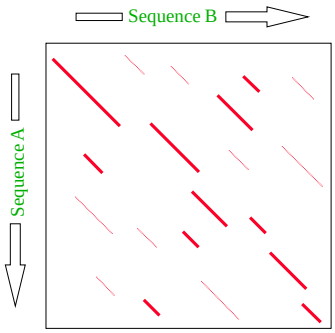
FASTA Algorithm



1

First run
(identities)

FASTA Algorithm



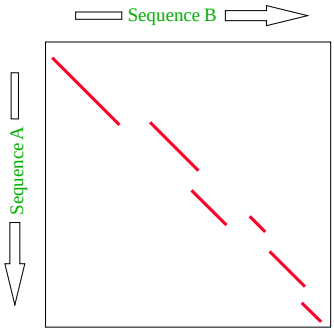
2

Rescoring using
PAM matrix

high score
low score

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

FASTA Algorithm

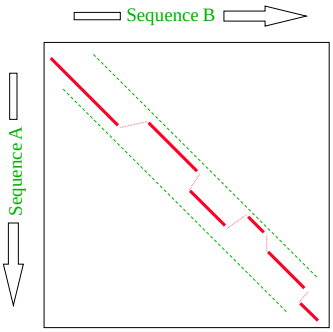


3

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 optimization 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

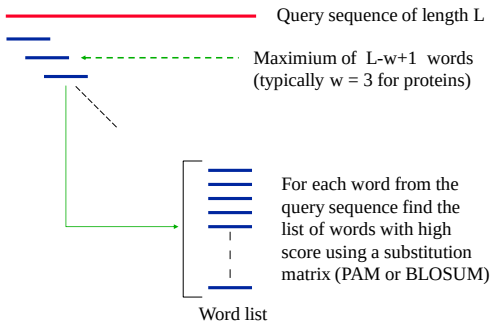
- When init1=init0=opt:
  100 % homology over the matched stretch.
- When initn > init1:
  more than 1 matching region in the database with poorly matching separating regions.
- When opt > initn:
  the matching regions are greatly improved by adding gaps in one or both of the sequences.

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.

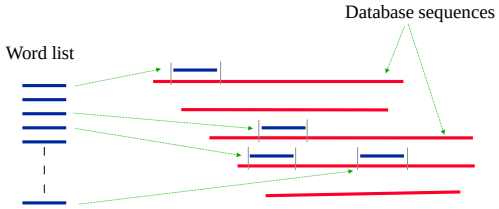
BLAST Algorithm

1



BLAST Algorithm

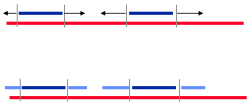
2



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