Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

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 acbcd and cbabd can be aligned in different ways

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

The String Alignment Problem

given: two strings CTCATG and TACTTG

score:
- C T C A T G
  |   |
- T A C T T G

3 · (2) + 3 · (-1) = 3

score:
- C T C A - T - G
  |   |   |   |
- T - A C T T G

4 · (2) + 4 · (-1) = 4
Entropy and Redundancy of Language

The sequences are 65% identical

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

Scoring Alignments

** CUR**** F*****w****** D***** DIS****AND P***
**BLES****FR*****B******BR*****AND ***** AG**

Substitution Matrices

- **Dayhoff (or MDM, or PAM)** -
  Derived from global alignments of closely related sequences
  PAM100 - number refers to evolutionary distance
  (Percentage of Acceptable point Mutations per \(10^8\) years)
  - 300 million years
  - 200 million years
  - 100 million years

- **BLOSUM (BLOcks SUbstitution Matrix)** -
  Derived from local, ungapped alignments of distantly related sequences
  BLOSUM62 - number refers to the minimum percent identity

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)**

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

THERE IS NO “ONE SIZE FITS ALL” MATRIX!

Search and alignment entropy

- **Information content per position:**
  - pam10: 3.43 bits
  - pam120: 0.98 bits
  - pam160: 0.70 bits
  - pam250: 0.38 bits
  - blossum62: 0.70 bits

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

Search and alignment entropy

<table>
<thead>
<tr>
<th>Query length</th>
<th>Substitution matrix</th>
<th>Gap costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>PAM-30</td>
<td>(9,1)</td>
</tr>
<tr>
<td>35–50</td>
<td>PAM-70</td>
<td>(10,1)</td>
</tr>
<tr>
<td>50–85</td>
<td>BLOSUM-80</td>
<td>(10,1)</td>
</tr>
<tr>
<td>&gt;85</td>
<td>BLOSUM-62</td>
<td>(11,1)</td>
</tr>
</tbody>
</table>

**FASTA Algorithm**

1. Sequence B
   - First run (identities)
   - Rescoring using PAM matrix
     - high score
     - low score

The score of the highest scoring initial region is saved as the **init score**.
### FASTA Algorithm

**Sequence A**

**Sequence B**

**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 **init score**.

**FASTA Algorithm**

**Sequence B**

**Sequence A**

**FASTA Algorithm**

**Alignment optimization using dynamic programming**

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

**FASTA Algorithm**

**Sequence B**

**Sequence A**

**FASTA Algorithm**

**Sequence B**

**Sequence A**

**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 Algorithm**

**Sequence B**

**Sequence A**

**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

**Query sequence of length L**

Maximum of **L-w+1** words (typically **w = 3** for proteins)

For each word from the query sequence find the list of words with high score using a substitution matrix (PAM or BLOSUM)

Word list
BLAST Algorithm

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

Word list

Database sequences

Gapped BLAST

- The Gapped Blast algorithm allows gaps to be introduced 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