

History of sequence searching

- 1970: Needleman-Wunsch
- 1980: Smith-Waterman
- 1985: FASTA
- 1990: BLAST
- 1997: BLAST2

Sequence alignment heuristics

Bioinformatics Fall 2011

1

BLAST

- Basic Local Alignment Search Tool
 - Altschul et al. 1990, 1997
- Heuristic for local alignment
- Designed specifically for database searches
- Idea: good alignments contain short lengths of high-scoring matches

Sequence alignment heuristics

Bioinformatics Fall 2011

2

Steps of BLAST

1. Filter low complexity regions (optional)
2. *Query words* of length 3 (for proteins) or 11 (for DNA) are created from query sequence using a sliding window

```
MEFPGLGSLGTSEPLPQFVDPALVSS
MEF
EFP
FPG
PGL
GLG
```

Assumption: small query, large database

Sequence alignment heuristics

Bioinformatics Fall 2011

3

Steps of BLAST

3. Score all *possible* words against a query word.
4. Select a *neighborhood word score threshold* (T) so that only most significant hits are kept. Approximately 50 hits per query word.
5. Repeat 3 and 4 for each query word in step 2. Total number of high scoring words is approximately $50 * (\text{query sequence length})$.

Sequence alignment heuristics

Bioinformatics Fall 2011

4

Steps of BLAST

6. Organize the high-scoring words into a lookup table
7. Scan database sequence for an exact match to high-scoring words. Each match is a *seed* for an ungapped alignment.

Sequence alignment heuristics

Bioinformatics Fall 2011

5

Steps of BLAST

8. (*Original BLAST*) extend matching words to the left and right using ungapped alignments. Extension continues as long as score increases or stays same. This is an HSP (high scoring pair).

(*BLAST2*) Matches along the same diagonal (think dot plot) within a distance A of each other are joined and then the longer sequence extended as before. Need at least two contiguous hits for extension. (Requires lower T). This is an HSP (high scoring pair).

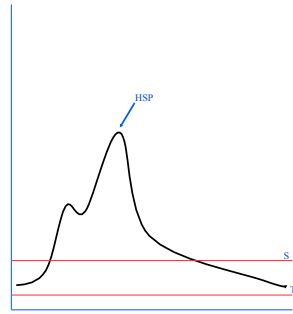
Sequence alignment heuristics

Bioinformatics Fall 2011

6

Steps of BLAST

9. Using a cutoff score S_c , keep only those HSPs that have a score at least S_c .
10. Determine the statistical significance of each remaining HSP.
11.
 - (Original BLAST) Only ungapped alignments; sometimes combined together
 - (BLAST2) Extend the HSPs using gapped alignment



Sequence alignment heuristics

Bioinformatics Fall 2011

7

Summarizing BLAST

- One of the few algorithms to make it as a verb
 - Blast(v): to run a BLAST search against a sequence database
- Extension is the most time-consuming step
- BLAST2 reported to be 3 times faster than the original version at same quality

Sequence alignment heuristics

Bioinformatics Fall 2011

8

BLAST variants

Program	Query	Database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide (six-frame translation)	Protein
TBLASTN	Protein	Nucleotide (six-frame translation)
TBLASTX	Nucleotide (six-frame translation)	Nucleotide (six-frame translation)

Sequence alignment heuristics

Bioinformatics Fall 2011

9

FASTA

- Another sequence alignment algorithm

Sequence alignment heuristics

Bioinformatics Fall 2011

10

Steps of FASTA

1. Find k-tups in the two sequences (k=1-2 for proteins, 4-6 for DNA sequences)
2. Select top 10 scoring "local diagonals" with matches and mismatches but no gaps.
 - a. For proteins, each k-tup found is scored using a score matrix such as PAM250
 - b. For DNA, use the number of k-tups found
 - c. Penalize intervening regions of mismatches

Sequence alignment heuristics

Bioinformatics Fall 2011

11

Finding k-tups

```

position 1 2 3 4 5 6 7 8 9 10 11
protein 1 n c s p t a . . . . .
protein 2 . . . . . a c s p r k
amino acid      position in      offset
                  protein A protein B   pos A - posB
-----
a                  6          6           0
c                  2          7           -5
k                  -          11           -
n                  1          -           -
p                  4          9           -5
r                  -          10           -
s                  3          8           -5
t                  5          -           -
    
```

Note the common offset for the 3 amino acids c,s and p.
A possible alignment is thus quickly found -

```

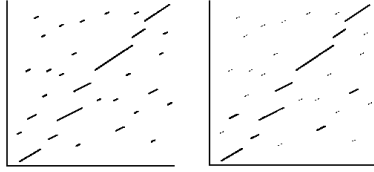
protein 1 n c s p t a
           | | |
protein 2 . a c s p r k
    
```

Sequence alignment heuristics

Bioinformatics Fall 2011

12

FASTA



Finding 10 best diagonal runs

Sequence alignment heuristics

Bioinformatics Fall 2011

13

FASTA

3. Rescan top 10 diagonals (representing alignments), and trim off the ends of the regions to achieve highest scores. The highest score at this point is called *init1*.
4. Join regions that are consistent with gapped alignments. (maximal weighted paths in a graph). The cumulative score is called *initn*.

Sequence alignment heuristics

Bioinformatics Fall 2011

14

FASTA

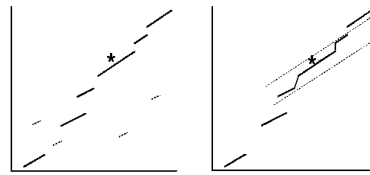
5. After finding the best initial region (step 3), FASTA performs a DP global alignment in a band centered on the best initial region, and uses the score as the optimized score *opt*.
for proteins, band size = 16 if *ktup* = 2
32 if *ktup* = 1
6. Compute statistics based on *init1*, *initn*, *opt*.

Sequence alignment heuristics

Bioinformatics Fall 2011

15

FASTA



init1, *initn*, and dynamic programming

Sequence alignment heuristics

Bioinformatics Fall 2011

16

Summarizing FASTA

- Statistics based on histograms on values of *init1*, *initn*, *opt*
- Begins with exact matches unlike BLAST
- Less of a statistical basis for comparison
- Quality and complexity similar to BLAST

Sequence alignment heuristics

Bioinformatics Fall 2011

17

Scalable tools

- Megablast
 - Uses a longer seed length of 28
- BLAT
 - BLAST indexes the query and scans the database
 - BLAT indexes the database, thus avoiding a linear scan of the database
 - Optimized for near-exact matches
 - 1 GB index for the human genome

Sequence alignment heuristics

Bioinformatics Fall 2011

18

Sensitive tools

- Patternhunter
 - Gapped seeds 111010010100110111
- PSI-BLAST
 - Construct position specific score matrix based on good hits and iterate
 - Relevance feedback

Sequence alignment heuristics

Bioinformatics Fall 2011

19

Position Specific Scoring Matrix (PSSM)

- Estimate the information content of each column of a multiple alignment.
- Can align against a PSSM
- PSSM can be constructed using multiple alignment and examining the characters aligned at the specific column.

Sequence alignment heuristics

Bioinformatics Fall 2011

20

Information in PSSMs

- Information theory: amount of information contained at each column.
- No information: amount of uncertainty can be measured as $\log_2 20 = 4.32$ for amino acids, since there are 20 amino acids. For nucleic acid sequences, the amount of uncertainty can be measured as $\log_2 4 = 2$.
- If a column is completely conserved then the uncertainty is 0 – there is only one choice.

Sequence alignment heuristics

Bioinformatics Fall 2011

21

Measure of Uncertainty

- Measured as the entropy
 - $H_c = - \sum_i p_{ic} \log p_{ic}$
 - p_{ic} is the probability of occurrence of symbol i in column c
 - $H = \sum_c H_c$

Sequence alignment heuristics

Bioinformatics Fall 2011

22

PSI-BLAST Algorithm

1. Perform initial alignment with BLAST using BLOSUM 62 substitution matrix
2. Construct a multiple alignment from matches
 1. E-value cutoff
3. Prepare a position specific scoring matrix
4. Use PSSM profile as the scoring matrix for a second BLAST run against database
5. Repeat steps 2-4 until convergence

Sequence alignment heuristics

Bioinformatics Fall 2011

23

PSSM construction

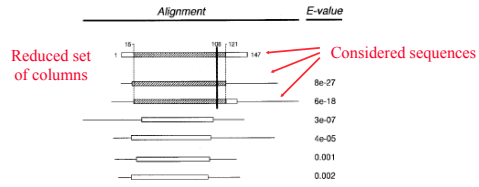
- Remove sequences identical to query
- Merge sequences more than 98% identical
- For any (non-gap) column c in query
 - R is the set of sequences with an aligned residue in c
 - Construct a reduced multiple alignment M_c
 - Consider only those columns that are represented in all of R
 - Compute weights for each sequence based on degree of similarity
 - Compute P_i , the background frequency of residue i over M_c
 - Compute weighted frequency f_i for each residue i .
 - Compute pseudocount g_i for each residue i (expectation based on score matrix).
 - Compute Q_i as the weighted sum of f_i and g_i .
 - Score for residue i is given by $\log(Q_i/P_i)$

Sequence alignment heuristics

Bioinformatics Fall 2011

24

PSSM construction



Sequence alignment heuristics

Bioinformatics Fall 2011

25

De novo sequence assembly

```

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

- Mouse: Over 30 million reads, total size over 30 GB
 - Only the two ends are read
- Finding overlaps is the time-consuming part
 - Mouse/PCAP: 560 out of 660 days, or 85% of the time
- Shortest common superstring problem (NP-hard)
- Hamiltonian path problem
- Quality indicators need to be used
- Repeat regions make the job harder
- TIGR, Phrap, VELVET, Arachne, Phusion

Sequence alignment heuristics

Bioinformatics Fall 2011

26

Mapping sequence assembly

- Align reads to an existing reference genome
 - Repeat regions make the job harder
- Understand variations
 - Copy number variation
 - Mutations
- Existing tools
 - Eland, Maq, SOAP, Bowtie

Sequence alignment heuristics

Bioinformatics Fall 2011

27