BLAST: Basic Local Alignment Search Tool

Chin Lung Lu

Computational Biology

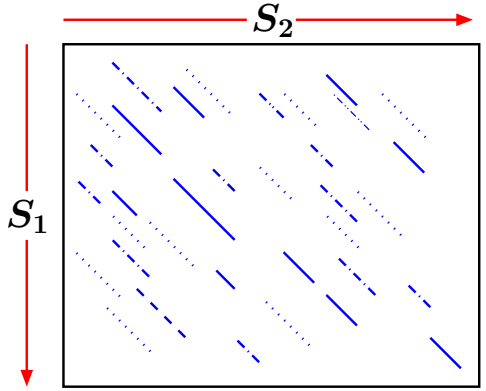
Analyses and Applications of Sequences

Classes of Comparison Algorithms

1. Dynamic programming algorithms: guarantee to find an optimal solution
   - Needleman-Wunsch algorihtm: global alignment (Needleman & Wunsch, 1970)
   - Smith-Waterman algorithm: local alignment (Smith & Waterman, 1981)
2. Heuristic algorithms: do not guarantee to find an optimal solution
   - FastA: (Pearson & Lipman, 1988)
   - BLAST: (Altschul et al., 1990)

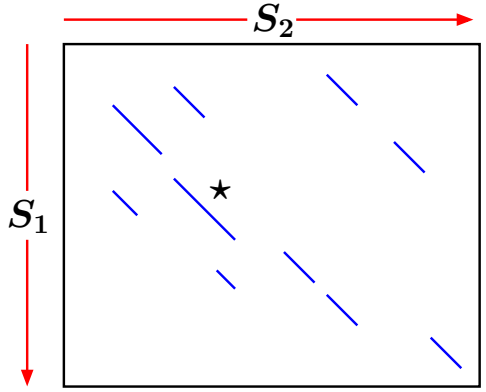
Algorithm of FastA

1. Identify k-length exactly matching hot-spots
   - hot-spot (k-tuple): a word of length k, where k =1–2 for protein and k =1–6 for DNA



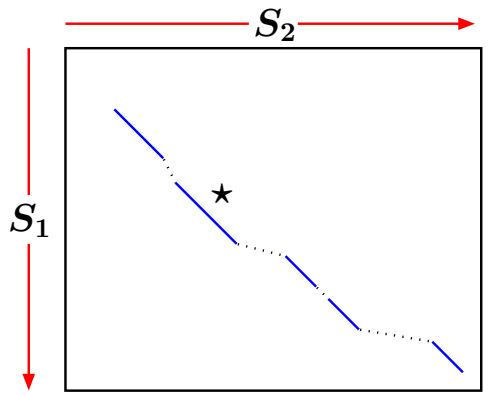
Algorithm of FastA

2. Rescore the 10 best scoring region using a scoring matrix and find $init_1$ socre:



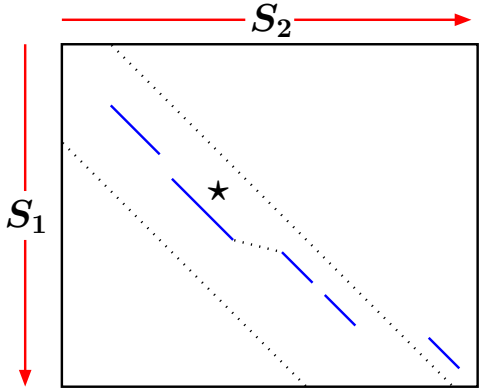
Algorithm of FastA

3. Find $init_N$ score by combining good alignments into a larger alignment allowing some spaces:



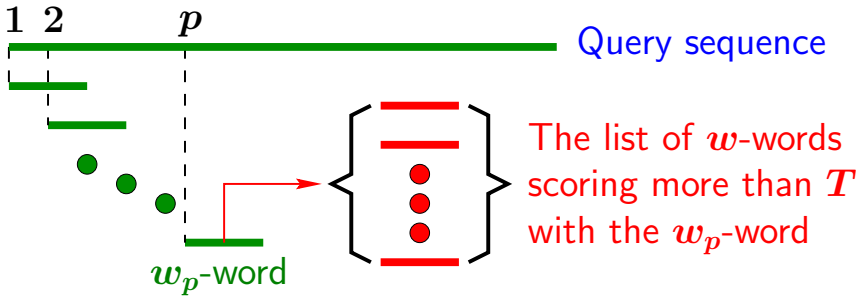
Algorithm of FastA

4. Find opt score by computing the optimal local alignment in the band consisting of 16 (or 32) diagonals around the diagonal with socre $init_1$:



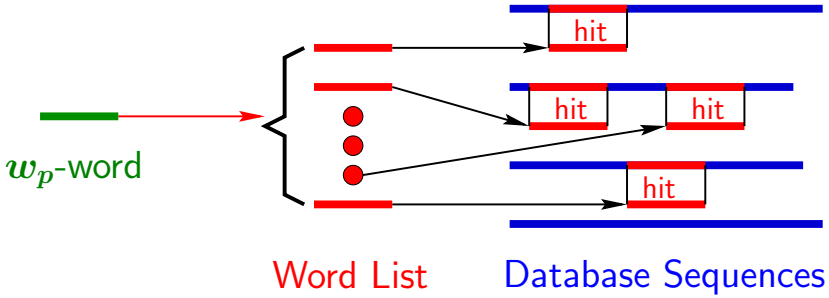
Algorithm of Ungapped BLAST

1. Preprocessing of query sequence: for each position p of the query, find the list of all w-words whose scores are greater than T when paired with the w-word starting at p (w_p-word) in the query



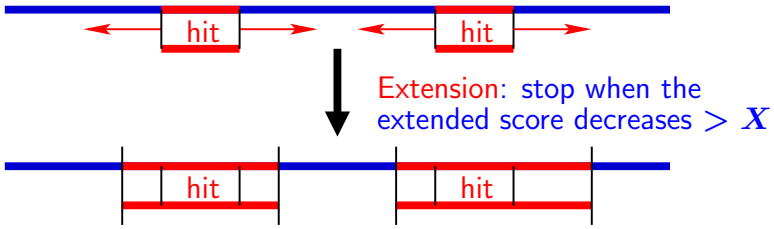
Algorithm of Ungapped BLAST

2. Generation of hits: for each word list, find all exact matches (hits) with the database sequences



Algorithm of Ungapped BLAST

3. Extension of hits: for each hit, extend ungapped alignment in both directions to find the alignments whose scores are greater than the threshold S



High scoring Segment Pair (HSP): the extended score $\geq S$

Central Idea of Ungapped BLAST

- A statistically significant alignment is likely to contain a high-scoring pair of aligned words.
  - Scan the database for words that score at least T when aligned with some word within the query sequence (such a word pair is called a hit).
- Check whether each hit lies within an alignment with score sufficient to be reported.
  - Extend a hit in both directions, until the running alignment's score dropped more than X below the maximum score yet attained (typically accounts for $> 90\%$ of execution time).

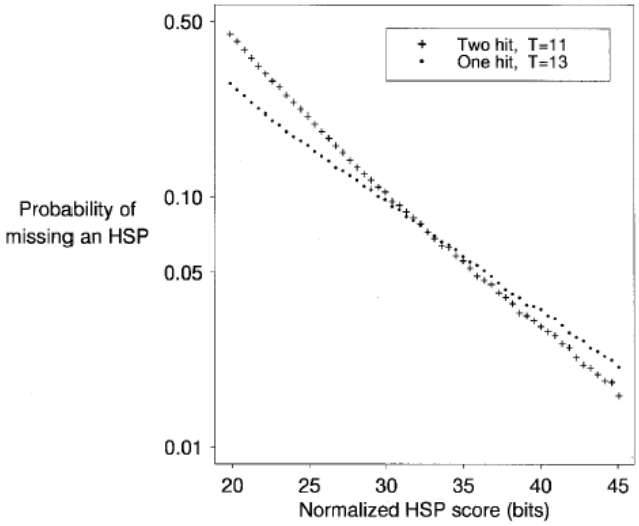
Two-Hit Method

- Since the extension step consumes most of the processing time, it is desirable to reduce the number of the performed extensions.
- How to reduce the number of extensions?
- **Observation:** an HSP of interest is much longer than a single word pair and may hence involve multiple hits on the same diagonal and within a relatively short distance of one another.
- Two-hit method: invoke an extension only when two non-overlapping hits are found within distance A of one another on the same diagonal

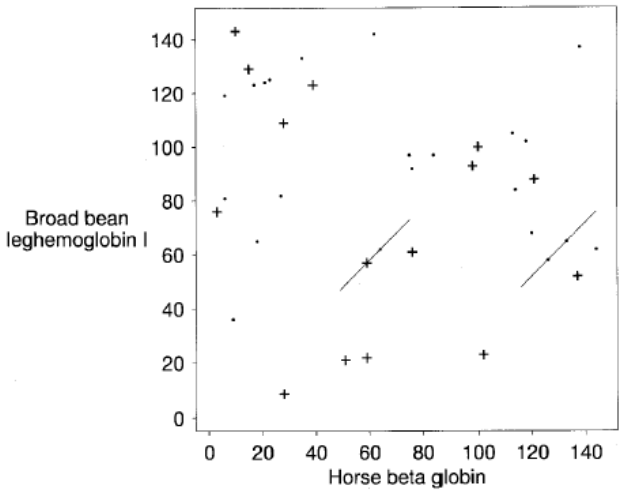
Speed Up by Two-Hit Method

- A higher value of T yields greater speed, but also yields an increased probability of missing weak similarities.
- In the two-hit method, we require two hits rather than one to invoke an extension and hence the threshold parameter T must be lowered to retain comparable sensitivity.
- As a result, many more single hits are found, but only a samll fraction have an associated second hit on the same diagonal that triggers an extension.

Sensitivity of Two- & One-Hits



Relative Speed of 2-Hit & 1-Hit



15 "+" for hits with score > 13 and 22 "·" for hits with score > 11

E-values

- In the sufficiently large sequence lengths m and n, the expected number of HSPs with score at least S (E-value for the score S) is $E = Kmne^{-\lambda S}$.
- Doubling m or n should double the number of HSPs attaining a given score.
- For an HSP to attain the score $2x$, it must attain the score x twice in a row, so one expects E to decrease exponentially with score.
- K and λ are the natural scales for the search space size and the scoring system respectively.

Bit Scores

- Raw scores have little meaning without detailed knowledge of the scoring system used.
- Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years.
- A raw score is normalized as bit score (S')

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

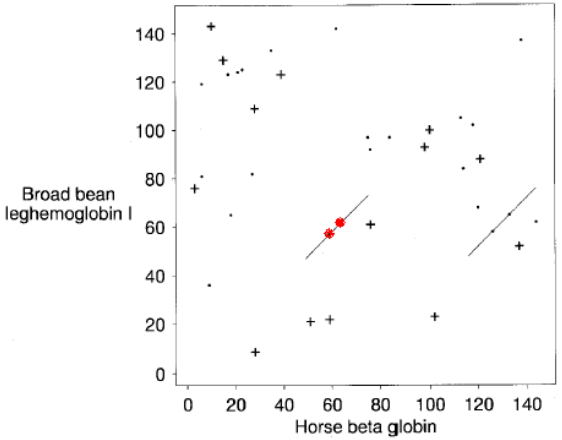
- Then the E-value corresponding to a given bit score S' is simply $E = mn2^{-S'}$.

Gapped Blast

1. Find two non-overlapping hits of score at least T, within a distance A of one another, to invoke an ungapped extension.
2. If the generated HSP has normalized score at least S_g bits, then a gapped extension is triggered.
3. The resulting gapped alignment is reported only if it has an E-value low enough to be of interest.

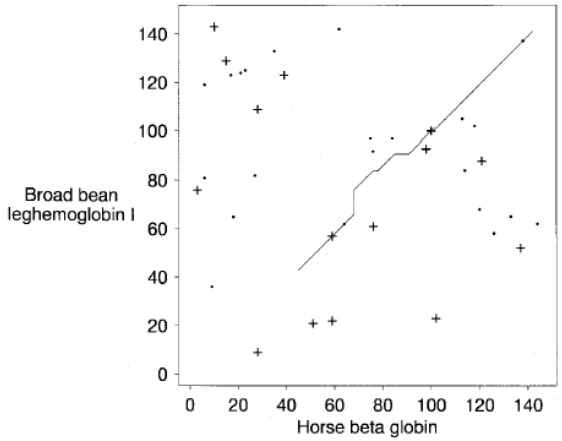
Example: Gapped Blast ❶

- The ungapped extension of the red hit pair obtains an HSP with score 23.6 bits ($T = 11, A = 40$).



Example: Gapped Blast ❷

- Then the gapped extension generates an alignment with score 32.4 bits and E-value of 0.5.



How Gapped Extension?

- Find a seed (a single aligned pair of residues)
  - Along the HSP, find the length-11 segment with highest alignment score.
  - Use its central residues pair as the seed.
- Starting from the seed, the dynamic programming proceeds both forward and backward through path graph.

Illustration: Gapped Extension