Sequence Alignment

Linear space and
Heuristic techniques

Global Alignment (review)

- Input: Two strings, labeled *s* and *t*
  - |*s*| is *n*
  - |*t*| is *m*
- Output: Two strings s_A and t_A such that:
  - s_A and t_A are of equal length L
  - Characters must be in same order, with "-" spacers as needed
  - If $s_A[i]$ = '-', then $t_A[i] \neq$ '-'
  - If $t_A[i]$ = '-', then $s_A[i] \neq$ '-'

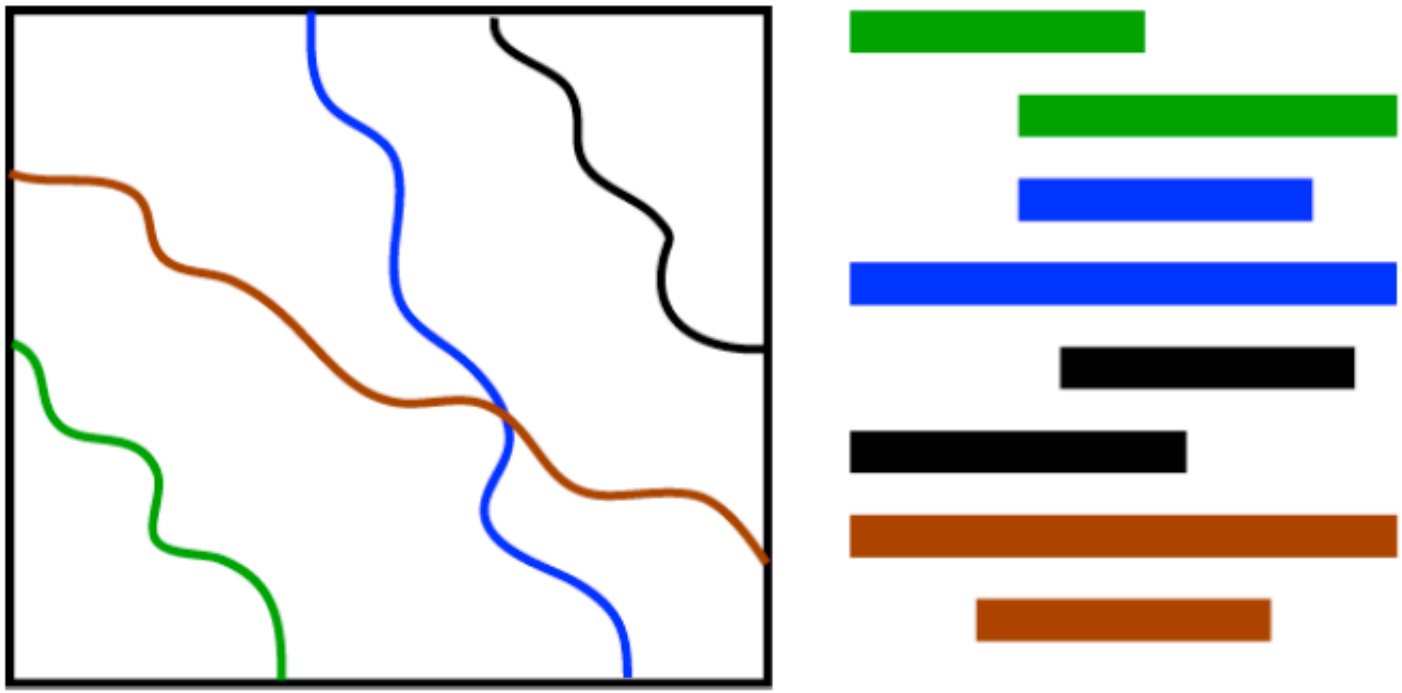
End-gap free alignment

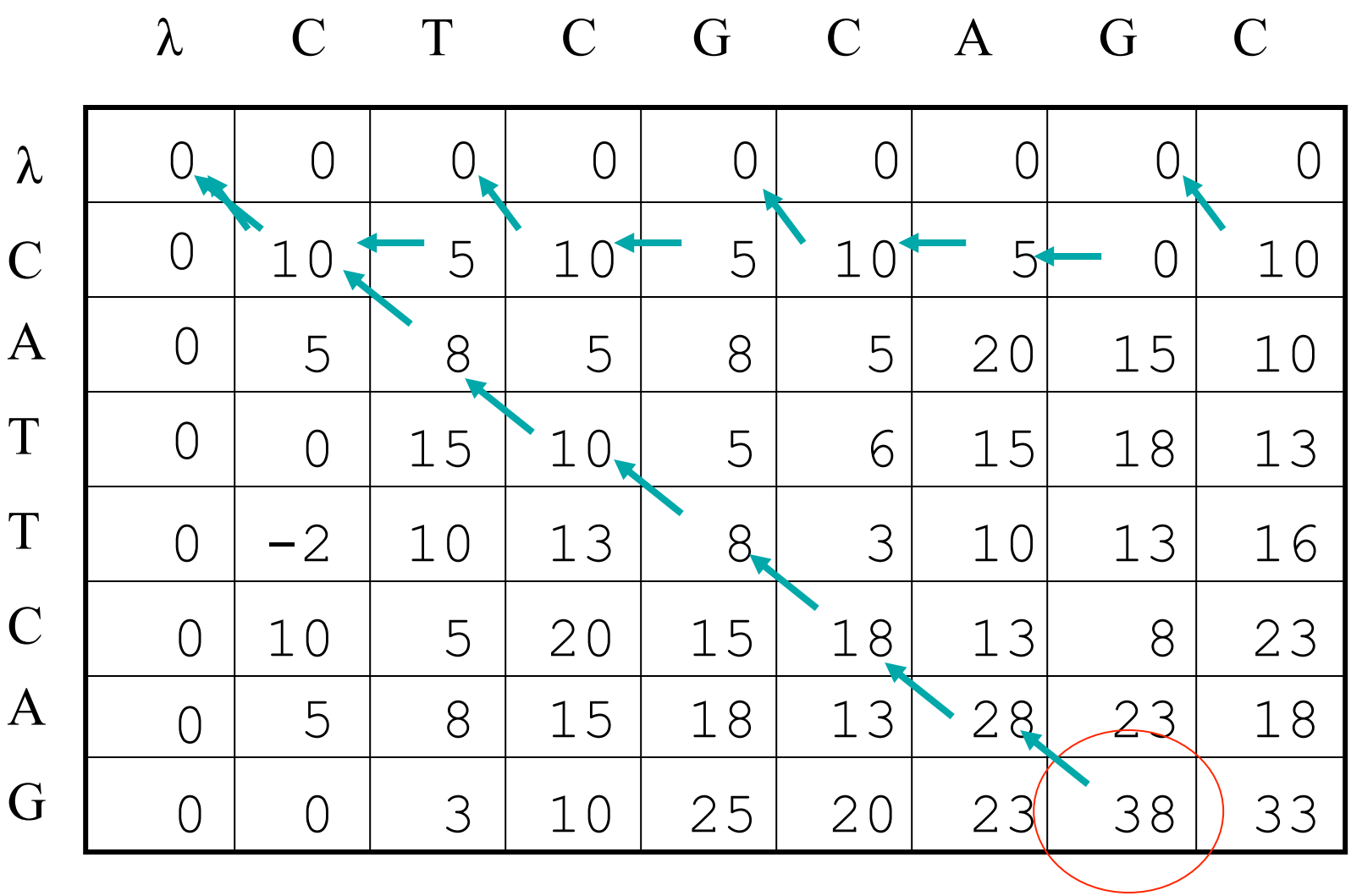
- We often don't want to penalize gaps at the start or end of the alignment, especially when comparing short and long sequences.

- Same as global alignment, except:
  - Initialize with zeros (free gaps at start)
  - Locate max in the last row/column (free gaps at end)

- Also called *semiglobal alignment.*

Example paths



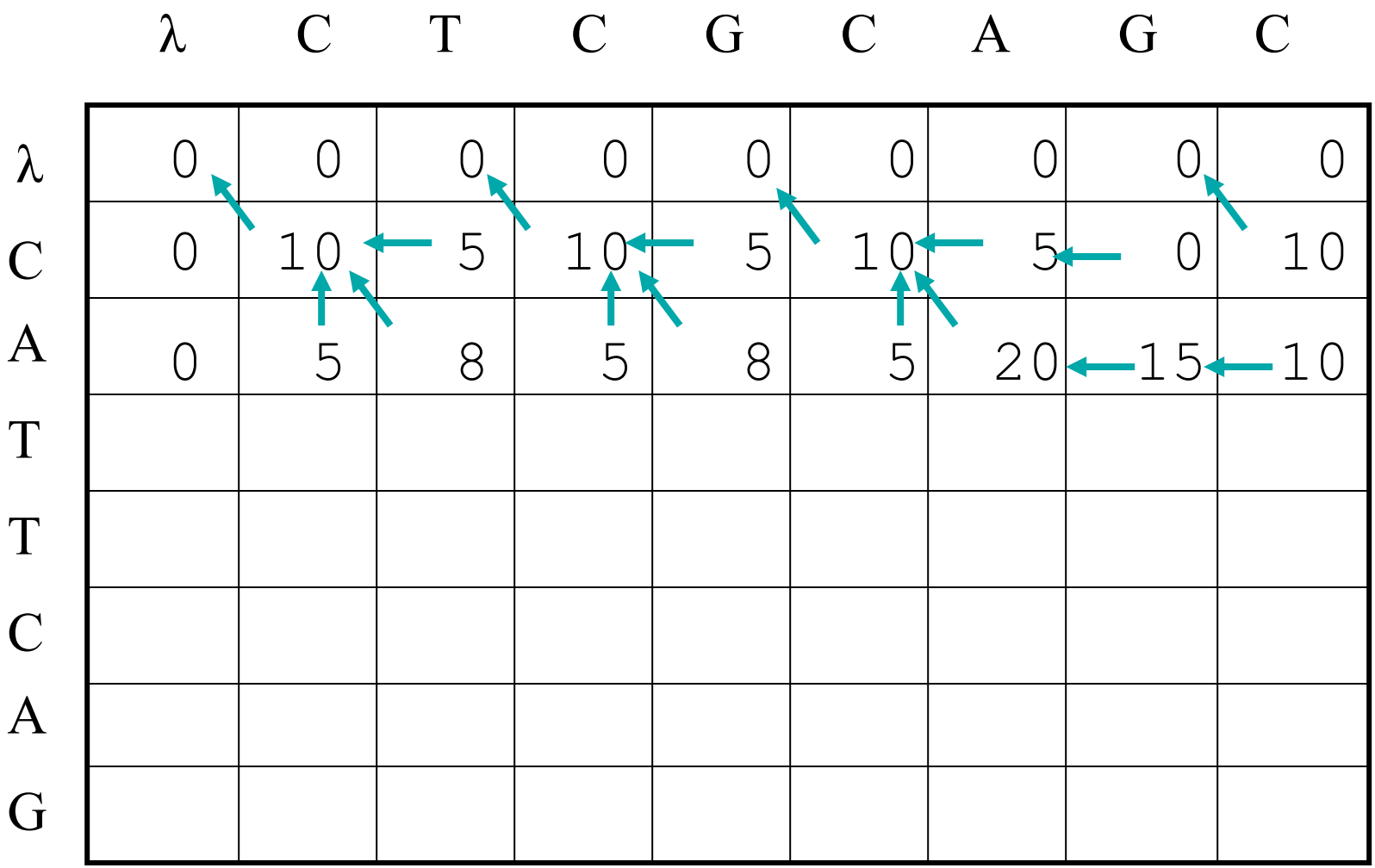


+10 for match, -2 for mismatch, -5 for space (rowwise)

Alignments in linear space

- The space required for the alignment table is O(*mn*). For large strings, this is not good.

- The Hirschberg technique reduces this to O(min(m,n)). We'll discuss this in greater depth today.



Starting IDEA: We only need the *previous* row to calculate the *next*

Assertion

- Before, we made the following observation:

$$A(n,m) = A(1..i,1..j) + A(i..n, j..m)$$

What if the optimal alignment goes through A(*i,j*)?
Does this help?

The trick

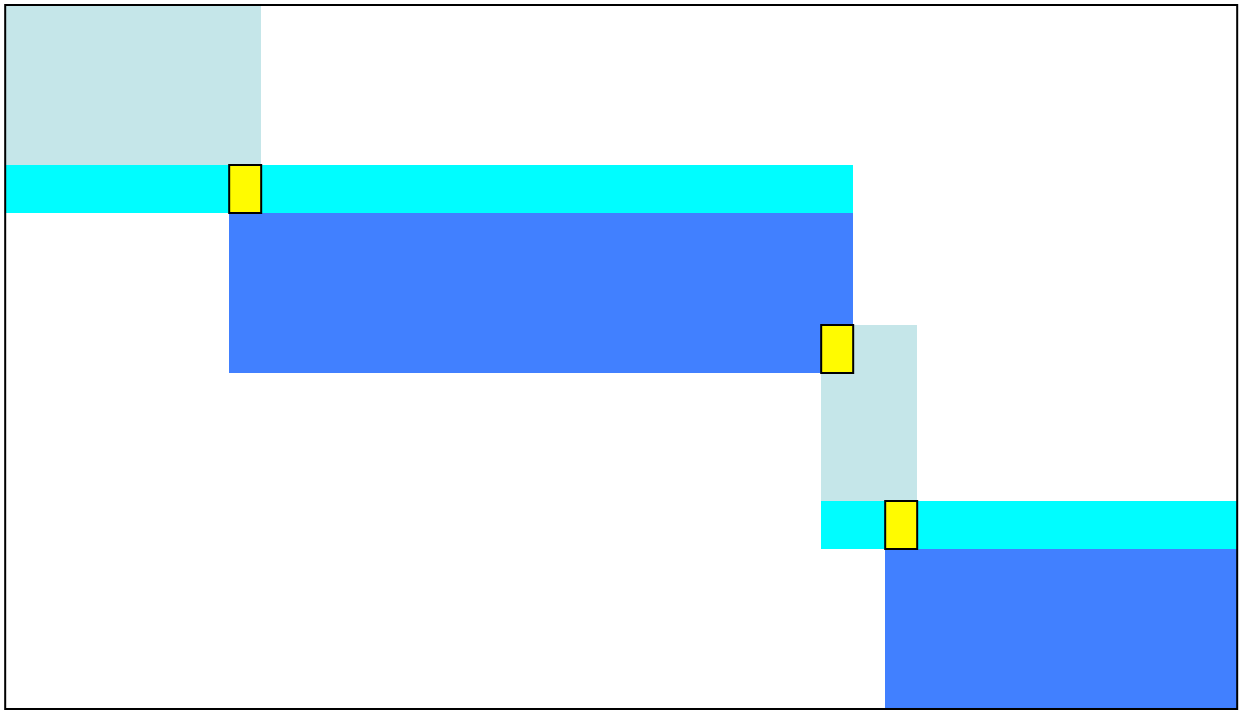
- Lemma:

$$A(n,m) = \max_{0 \le k \le m} \left\{ A\left(\frac{n}{2}, k\right) + A^r\left(\frac{n}{2}, m-k\right) \right\}$$

Where A^r is the reverse of strings *s* and *t*

In other words, *s*[*n*]s[*n*-1]…*s*[1] aligned to *t*[*m*]t[*n*-1]…*t*[1]

Linear-space Alignments



*mn + ½ mn + ¼ mn + 1/8 mn + 1/16 mn + … = (2 mn) * 2*

Problem

- Prof. Hellmann in Biological Sciences has 5500 butterfly genes. She would like to see if they match any sequences in a BIG database.

- Armed with sequence alignment, what do want to do next?

Reality

- We could in theory perform a global alignment of each butterfly gene to the database.
- However, it turns out this will not be practical for most purposes.
- BLAST is one approach that will approximate the alignment, often orders of magnitude faster.

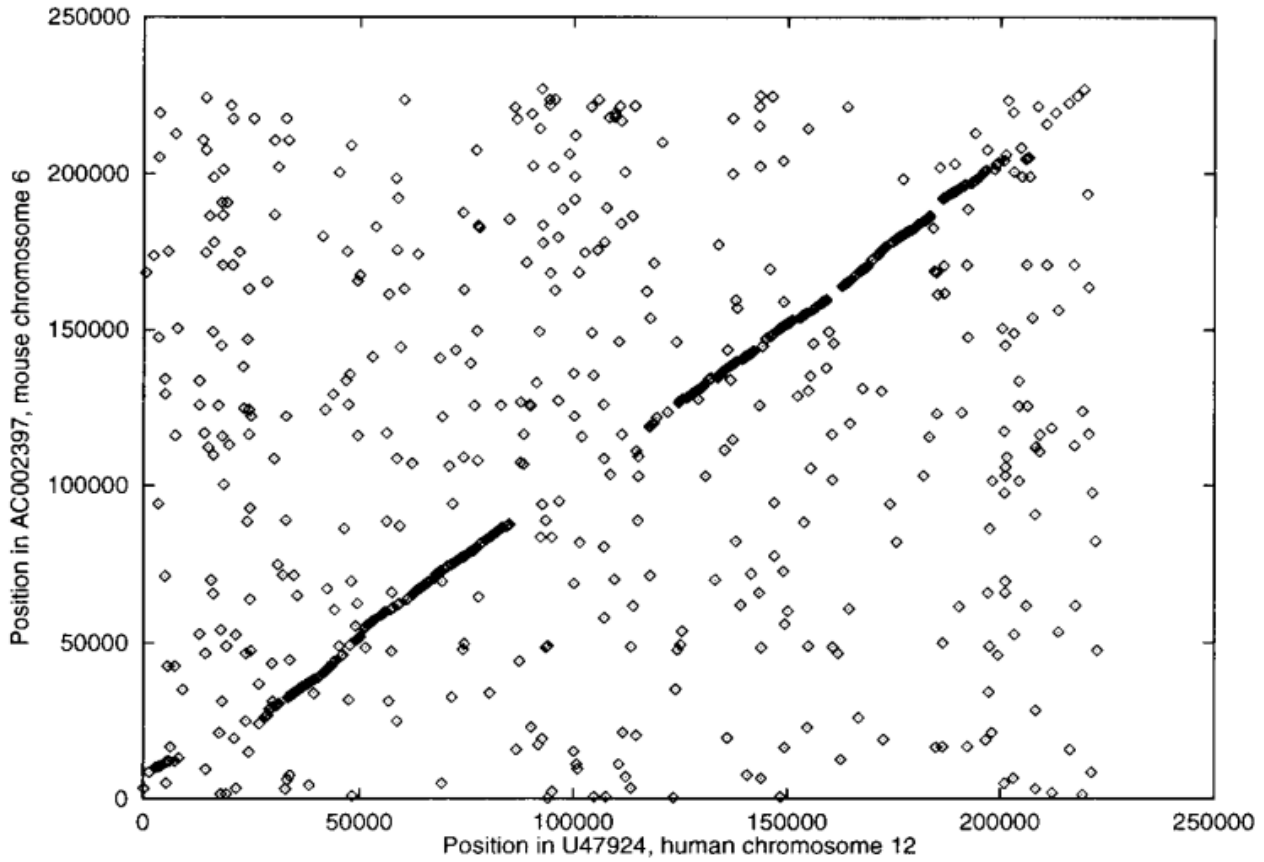
Simple dotplots

- Lets construct simple dot plots for the following strings:
  - s = xxixxxixixxgngab
  - t = bagngxxixixxxixx
- To make a dotpot:
  - Fill in a box/pixel if $x_i = x_j$, leave it empty if x_i != y_j
- See anything interesting?

Less simple dotplots

- We will construct dot plots for the same strings as before:
  - s = xxixxxixixxgngab
  - t = bagngxxixixxxixx
- To make the dotplot
  - Fill in a box/pixel at i,j if $x_i x_{i+1} x_{i+2} = x_j x_{j+1} x_{j+2}$
  - leave it empty otherwise
- See anything different?

Example Human-Mouse Dot Plot



The contenders

- FASTA - Fast-All
  - Pearson and Lipman (1988)
- BLAST
  - Altschul et al. (1990)

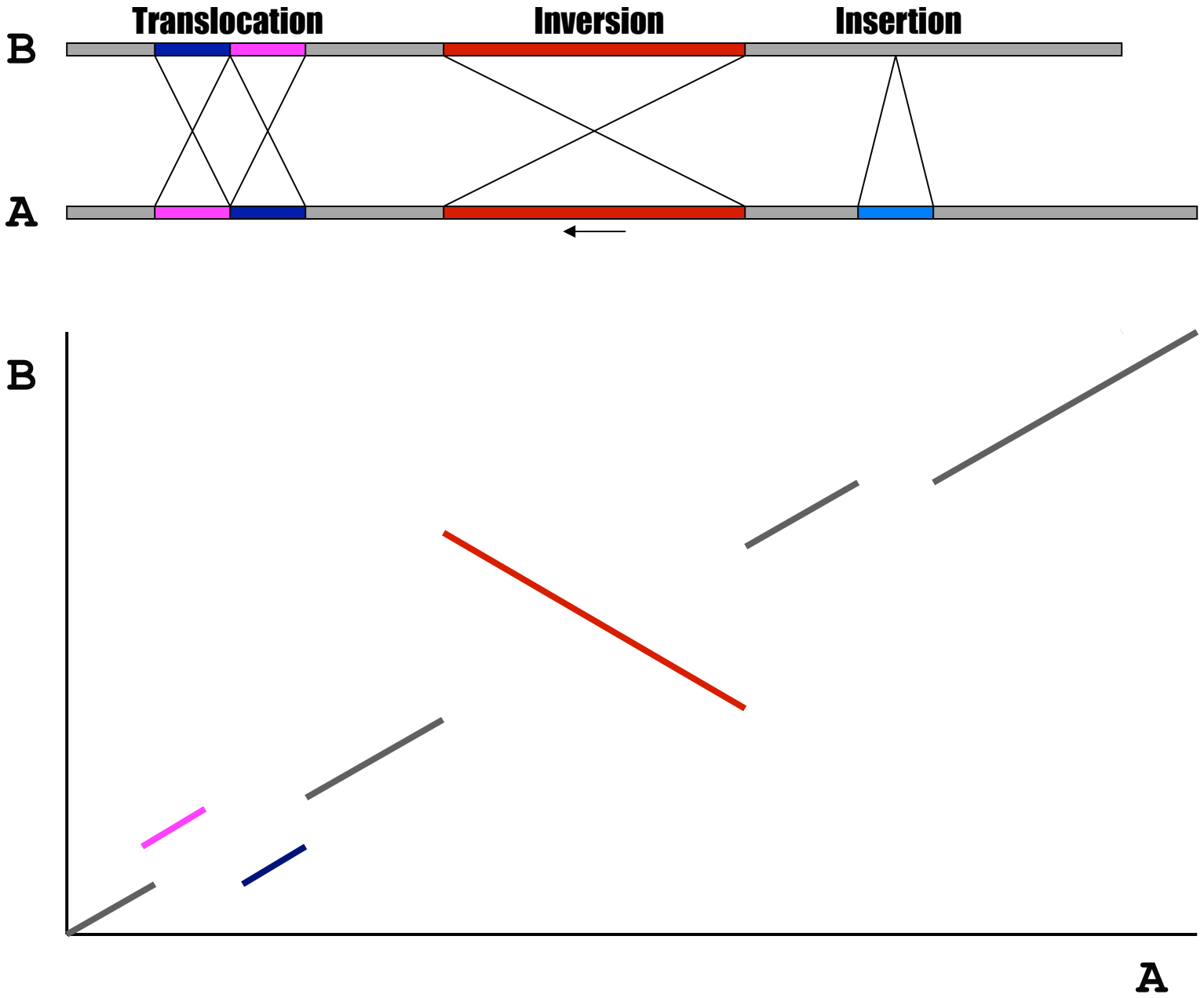
BLAST

- Basic Local Alignment Search Tool
- Used to quickly compare a protein or DNA sequence to a database.
- Output:
  - Alignment
  - Score
  - Significance (e-value)

“There is no such thing as a free lunch”

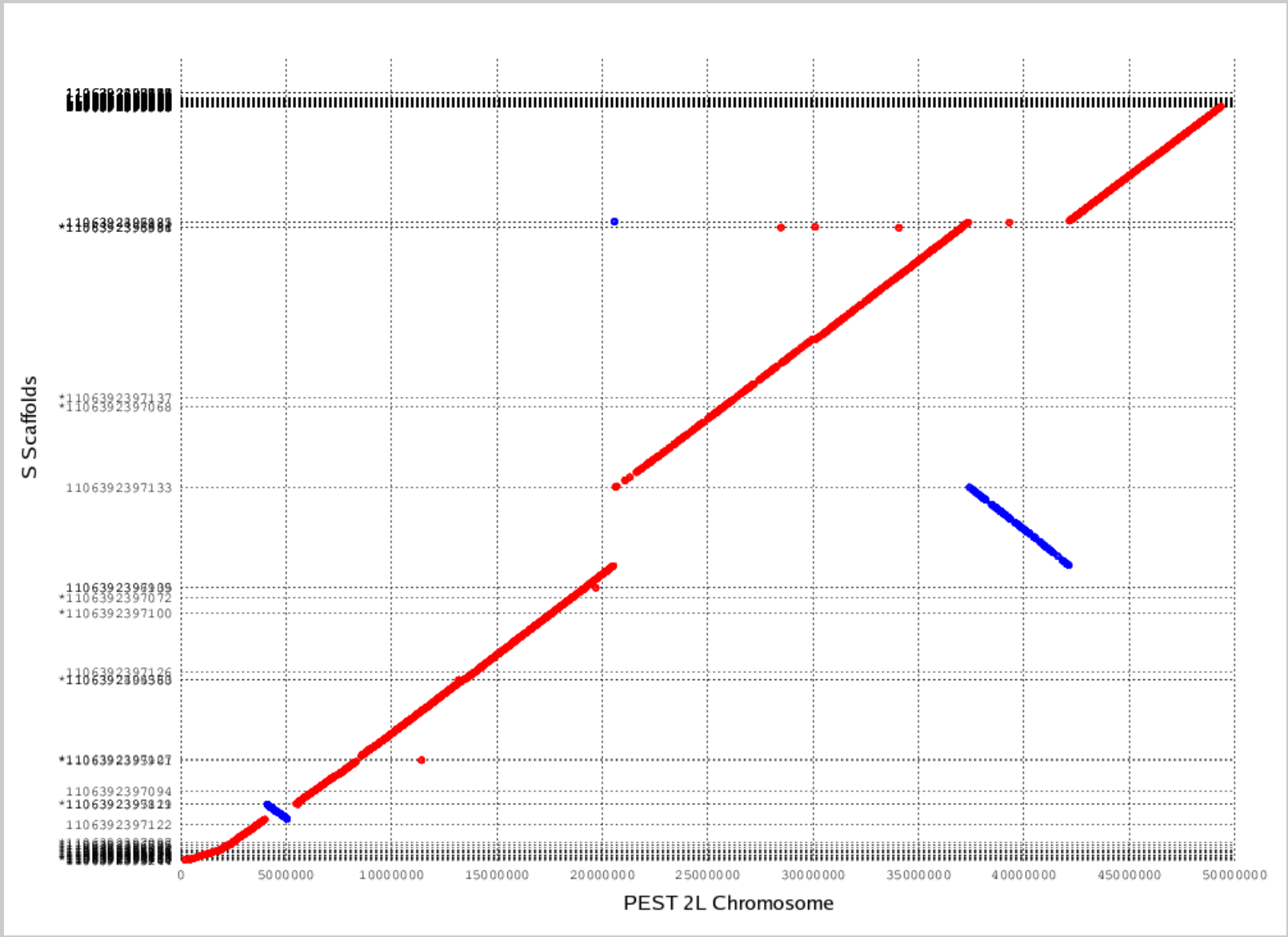
- BLAST is fast and highly sensitive compared to competitors.

- Disadvantages:
  - Misses some homologous matches
  - Alignment is not guaranteed to be optimal



http://mummer.sourceforge.net/manual/AlignmentTypes.pdf

Slide by Adam Phillppy


S Scaffolds
1106392397133
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000
40000000
45000000
50000000
PEST 2L Chromosome

Dirty details

- Four basic steps are performed in BLAST:
  - Compile a list of “interesting” words
  - Scan for these words, generating “hits”
  - Extend the hits into longer alignments
  - Determine if the longer hits are “good”

Compiling words

- Given a word length l = 4

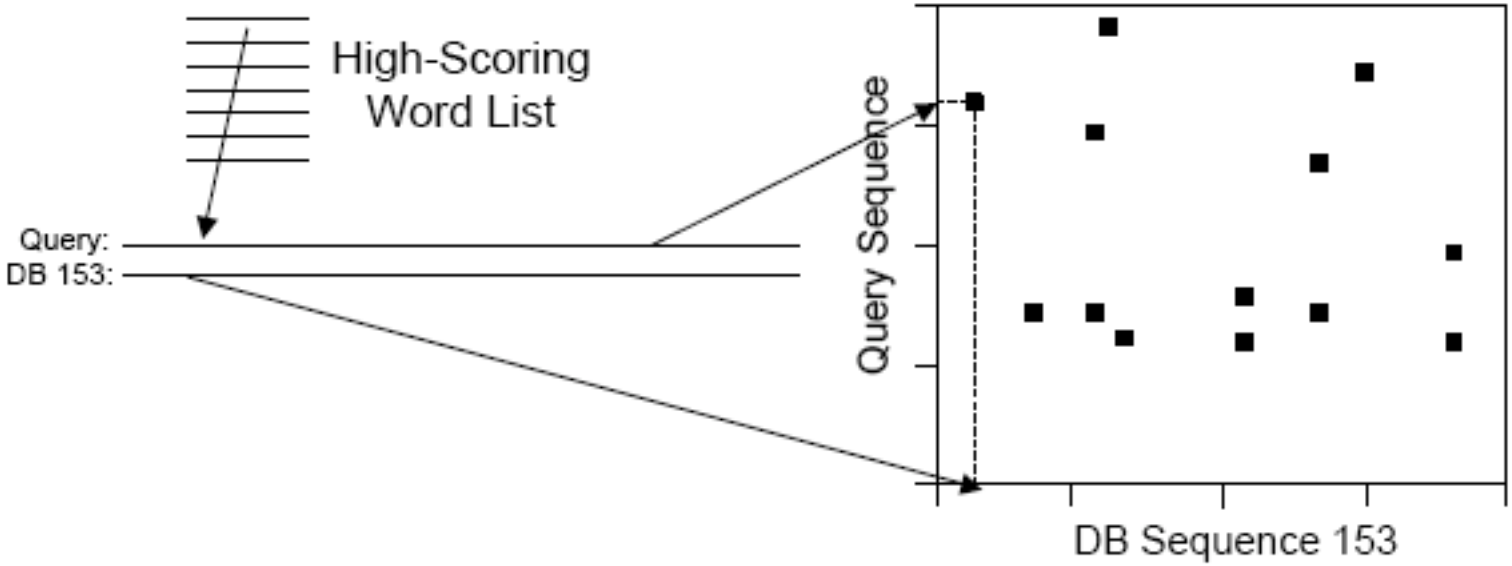
ATGCTGTTTTGGGAATGTGTG

ATGC
  TGCT
    GCTG
      CTGT
        ……
                TGTG

Collecting hits

- Scan the first word against the entire database.
- For every matching word, record a diagonal in a special table
- Repeat for the entire query.

Illustration



Finding runs and extending them (old version)

- Extend alignments “greedily” off each end; stop when the score drops below a threshold
- Alignments are ungapped, so extension is straightforward.
- All hits whose score is greater than a minimum score S are displayed.

Gapped BLAST

- Require two hits on the same “diagonal”
- Hits must be less than a specified distance away (antidiagonal difference). Alignment is explored in between matches (banded variant)
- Substantially reduces the number of extensions, and is a better heuristic in terms of biological value

E-value

- So what? We found a bunch of probably useless alignments, right?
- There are sophisticated statistics at the core of BLAST that produce an “e-value”, which is the probability of observing an alignment by chance.
- Based on work of Altshul and Karlin, based on extreme value statistics.