Comparative bacterial genomics

J. C. Setubal
VBI&CS
March 2010

Replicon sequence comparisons

- Basic tool: MUMmer
  - Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 2002 Jun 1;30(11):2478-83.
  - Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5(2):R12
- http://mummer.sourceforge.net

Basics of MUMmer

- It finds Maximal Unique Matches
- These are exact matches above a user-specified threshold that are unique
- Exact matches found are extended
- Data structure: suffix tree
  - Difficult to build but very fast
- Nucmer and promer
  - Both very fast

sample nucmer output (coords file)

```
/home/setubal/agro/comp/mummer/../../rhizogenes/v1/ctgs.fasta
/home/setubal/agro/comp/mummer/../../vitis/v3/all.fasta
NUCMER

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  | [TAGS]
==========================================================================================
   73024    73193  |   242351   242181  |      170       171  |    93.60  | Contig789       Contig608
     220     6244  |    38759    32766  |     6025      5994  |    86.64  | Contig791       Contig604
    2798     6297  |   174039   177532  |     3500      3494  |    83.31  | Contig791       Contig606
    3828     6297  |   124183   126645  |     2470      2463  |    81.80  | Contig791       Contig606
    4767     5392  |   551684   551059  |      626       626  |    82.11  | Contig791       Contig607
    8214     8453  |    30747    30508  |      240       240  |    84.65  | Contig791       Contig604
   15408    15987  |   181050   181624  |      580       575  |    86.23  | Contig791       Contig606
   63864    74254  |   191954   181567  |    10391     10388  |    89.08  | Contig791       Contig604
   77203    79534  |   178882   176555  |     2332      2328  |    84.35  | Contig791       Contig604
  157451   158456  |   139804   140812  |     1006      1009  |    82.09  | Contig791       Contig606
  157483   157800  |    58429    58110  |      318       320  |    89.13  | Contig791       Contig604
  163575   166223  |    62781    60133  |     2649      2649  |    78.80  | Contig791       Contig605
  166754   168442  |    49403    47716  |     1689      1688  |    85.79  | Contig791       Contig604
  171247   173701  |    45005    42556  |     2455      2450  |    88.17  | Contig791       Contig604
  171261   172115  |   157617   158476  |      855       860  |    86.30  | Contig791       Contig606
  181828   184458  |    41748    39140  |     2631      2609  |    93.13  | Contig791       Contig604
  184829   185852  |    38838    37821  |     1024      1018  |    91.61  | Contig791       Contig604
```

Homework statement

1. Pick any two fully sequenced bacterial genomes
   1. Suggestion: rhizobiales, pseudomonadaceae, or xanthomonadaceae (different species! same genus ok)
   2. NCBI taxonomy

   http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/
2. Download the primary chromosomes

   NCBI ftp site

   ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ file .fna
3. Shift and reverse complement one of the sequences as necessary
4. Download and install MUMmer
5. Run nucmer on your chromosomes
   1. Depending on results you may have to choose a different pair of sequences
6. Run your LIS program
7. Report the results

Longest Increasing Subsequence

- S = $x_1, x_2, \ldots, x_n$, distinct integers
- Increasing Subsequence of S is a subsequence $x_{i1}, x_{i2}, \ldots, x_{ik}$,

with $i_1 < i_2 < \ldots < i_k$ such that for all $1 \leq j < k$ we have $x_{ij} < x_{ij+1}$

- Longest Increasing Subsequence of S is an IS of maximum length

Based in part on U. Manber, *Introduction to Algorithms,* p.167

Example

9 44 32 12 7 42 34 92 35 37 41 8 20 27
83 64 61 28 39 93 29

9 44 32 12 7 42 34 92 35 37 41 8 20 27
83 64 61 28 39 93 29

Length = 8

There may be more than one solution

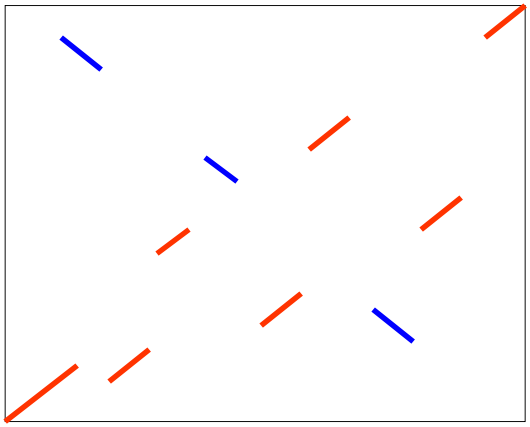
LIS algorithm

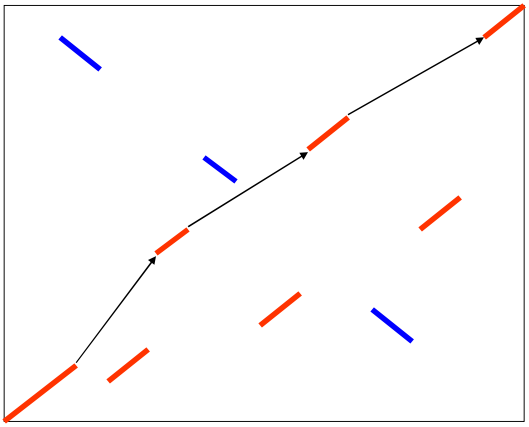
- Dynamic programming; let's use induction
- Base case: sequence with just one element
  - Trivial
- Induction hypothesis
  - We know the LIS for sequences of length < m
  - For each x_i, we know the LIS length that has x_i as last element LIS(x_i); the LIS value for the whole sequence up to m is max LIS(x_i) for all i
- Induction step
  - To find the solution including x_m find an x_i such that $i < m$, $x_i < x_m$, LIS(x_i) is max
  - LIS(x_m) = LIS(x_i) + 1

Complexity analysis

- As sketched algorithm is O(n^2)
  - For each x_i we have to scan all elements up to x_i
- It can be made to run in O(n log n) by doing binary search on the sub LIS values; requires separate data structure

LIS modeling of MUMmer alignment





Note that segments *of the same kind* need to be connected

Report format

1. Identify chromosomes chosen (accession (NC) number, organism name, chromosome size, whether sequence was shifted/RC'd)
2. Enclose copy of LIS program; should have inline clear documentation!
3. Report (txt OK)
   a. Nucmer coords file
   b. Explain how segments were used to determine input to LIS program
   c. Nucmer segments chosen by the LIS program
   d. length of LIS found in bp (add up all segment sizes chosen)
   e. Any comments you may have about the homework
4. Send all files in e-mail attachments to setubal@vbi.vt.edu