Biol5705
Module: Gene Sequence Analysis

Lecture 2
Homology Searching

Dr. Morgan Langille

Outline

- PSSMs/PSI-BLAST
- HMMs/HMMer
- RNA alignments
- Genome Alignments
- Assemblers
- Mappers

Different tools for homology searching

- Searching for protein families
- Aligning genomes
- Looking for RNA genes
- Combining overlapping sequences (assemblers)
- Finding the position of a sequence in a genome

One tool does not do it all

- Blast may give you an answer
  - BUT you could find the answer much quicker or with more precision by using the right tool!

PSI-BLAST

- Position Specific Iterated – BLAST
- A cycling/iterative method
  - Gives increased sensitivity for detecting distantly related proteins
  - Can give insight into functional relationships
  - Very refined statistical methods
- Fast – still based on BLAST methods
- Simple to use

PSI-BLAST

- Essentially we are using intermediate sequences to infer similarity between two sequences that are too dissimilar to link directly.

Profiles & PSSMs Need Multi-sequence Alignment

```
          1                                                  50
P43871-1 .......... ...IKKLDSN SIHAIISDIP YGIDYDDWDI LHSNTNSALG
S18997-1 LMSKIYQMDA VDWLKTLENC SVDLFITDPP YESL.EKYRQ IGTTTRLKES
P23192-1 EINKIHQMNC FDFLDQVENK SVQLAVIDPP YNL....... ..........
P29538-1 MDQRLICSNA IKALKNLEEN SIDLIITDPP YNLG.KDY.. ..........
P14751-1 TRHVYDVCDC LDTLAKLPDD SVQLIICDPP YNI....... ..........
P34721-1 KNFNIYQGNC IDFMSHFQDN SIDMIFADPP YFLS.NDG.L TFKNSIIQ..
P50178-1 ENAILVHADS FKLLEKIKPE SMDMIFADPP YFLS.NGG.M SNSGGQIV..
P20590-1 FLNTILKGDC IEKLKTIPNE SIDLIFADPP YFMQ.TEGKL LRTNGDEF..
S43876-1 GPETIIHGDC IEQMNALPEK SVDLIFADPP YNLQ.LGGDL LRPDNSKV..
P28638-1 EAKTIIHGDA LAELKKIPAE SVDLIFADPP YNIG.KNF.. ..........
P23941-1 DLGKLYNGDC LELFKQVPDE NVDTIFADPP FNLD.KEY.. ..........
P14230-1 RSCKIIVGDA REAVQGLDSE IFDCVVTSPP YWGL.RDY.. ..........
P14243-1 NGATLFEGDA LSVLRRLPSG SVRCIVTSPP YWGL.RDY.. ..........
Q04845-1 LNNMLLQGNC AETLKKLPDE SVNLVFTSPP YY....... ..........
S53866-1 WVNDIHEGDA EEVLAELPES SVHMVMTSPP YFGL.RDY.. ..........
P29568-1 .......... ...MNELKDK SINLVVTSPP YPMV.EIWDR LFSELNPKIE

Signature Sequence:                    DPP Y
```

How does PSI-BLAST work?

1) First, a standard blastp is performed

2) The highest scoring hits are used to generate a multiple alignment

3) A Position Specific Scoring Matrix (PSSM) is generated from the multiple alignment.

- Highly conserved residues get high scores
- Less conserved residues get lower scores
- The PSSM describes the sequence similarity between your query and all significant blastp hits

4) Another similarity search is performed, this time using the new PSSM as the query sequence.

- This PSSM (scoring matrix) is now customized to find sequences that are related to your original query
- Steps 2-4 can be repeated until convergence
- Convergence occurs when no new sequences appear after iteration

PSI-BLAST Example

Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

NEW - alignment score below the threshold on the previous iteration

- alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max 500 Go

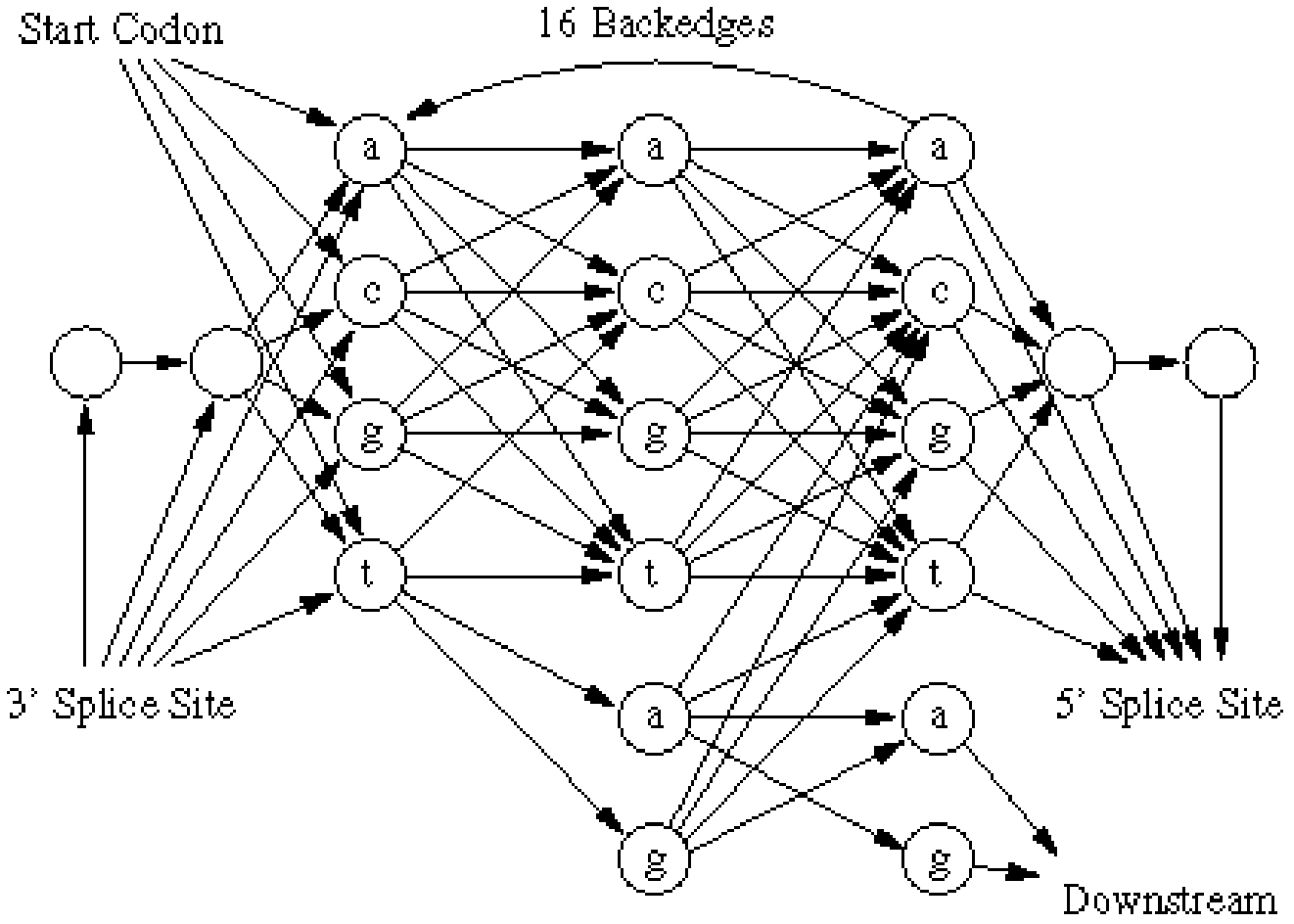
Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW NP_625085.1	tetracycline resistance protein [Streptomyces coelicolor A3(2)]	1172	1172	100%	0.0	98%	G
NEW ZP_07295505.1	tetracycline resistance protein TetP [Streptomyces hygroscopicus ATCC	650	650	97%	0.0	61%	
NEW YP_003342575.1	unnamed protein product [Streptosporangium roseum DSM 43021]	637	637	97%	0.0	58%	G
NEW YP_004811081.1	small GTP-binding protein [Streptomyces violaceusniger Tu 4113]	634	634	97%	0.0	58%	G
NEW ZP_06532924.1	tetracycline resistance protein [Streptomyces lividans TK24]	587	587	48%	0.0	99%	
NEW YP_004920166.1	unnamed protein product [Streptomyces cattleya NRRL 8057]	586	586	97%	0.0	58%	G
NEW YP_003118884.1	small GTP-binding protein [Catenulispora acidiphila DSM 44928]	560	560	97%	0.0	54%	G
NEW YP_003486610.1	unnamed protein product [Streptomyces scabiei 87.22]	472	472	95%	2e-156	50%	G
NEW ZP_06921723.1	translation elongation factor G [Streptomyces sviceus ATCC 29083]	468	468	95%	3e-155	50%	
NEW YP_003383525.1	small GTP-binding protein [Kribbella flavida DSM 17836]	467	467	97%	4e-155	48%	G
NEW ZP_04163174.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus mycoides Roc	462	462	96%	4e-153	38%	
NEW ZP_07276632.1	translation elongation factor G [Streptomyces sp. AA4]	462	462	97%	5e-153	47%	
NEW ZP_04151746.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus pseudomycoi	459	459	96%	9e-152	38%	
NEW ZP_04157524.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus mycoides Roc	457	457	96%	5e-151	38%	
NEW ZP_04198058.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus cereus AH60	457	457	96%	6e-151	38%	
NEW ZP_09405061.1	putative tetracycline resistance protein [Streptomyces sp. W007]	454	454	97%	1e-149	50%	
NEW ZP_04097136.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus thuringiensis	453	453	97%	1e-149	38%	
NEW ZP_04228518.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus cereus Rock3	451	451	96%	1e-148	38%	
NEW ZP_07308418.1	translation elongation factor G [Streptomyces viridochromogenes DSM	449	449	97%	6e-148	47%	

HMMs & HMMer

- The more powerful way to search for protein families than PSSMs

Hidden Markov Models



Hidden Markov Models in Bioinformatics

- **Used extensively in gene prediction**
- **Used to create Sequence Profiles and to classify sequences into families**
- **Used in Multiple Sequence Alignment**

Hmmer

- Suite of sequence analysis programs based on HMMs
- Used to build the Pfam database
- Available for free download at
  - http: hmmer.wustl.edu/

HMMER 3

- HMMER 2 was used for many years
  - Biggest draw back was always speed

- HMMER 3 released in 2011
  - Very fast with comparable speeds to BLAST
  - 100X faster than v2

HMMER Programs

- **hmmbuild** – build a HMM from multiple sequence alignment
- **hmmscan** – searches a query sequence(s) against a database of HMMs (used by PFAM)
- **hmmsearch**– searches a query HMM against a database of sequences (e.g. like psi-blast)
- **phmmer** – search a protein sequence vs a sequence database (e.g. like blastp)

HMMER Search & Software

- http://hmmer.janelia.org/search

- PFAM
  - http://pfam.sanger.ac.uk/

RNA Alignments

- RNA alignments are “special”
- RNA genes often have secondary structures that allow improved searching
- Improved searching is needed since
  - Must search in DNA space (less complex then protein sequences)
  - Often shorter length then proteins

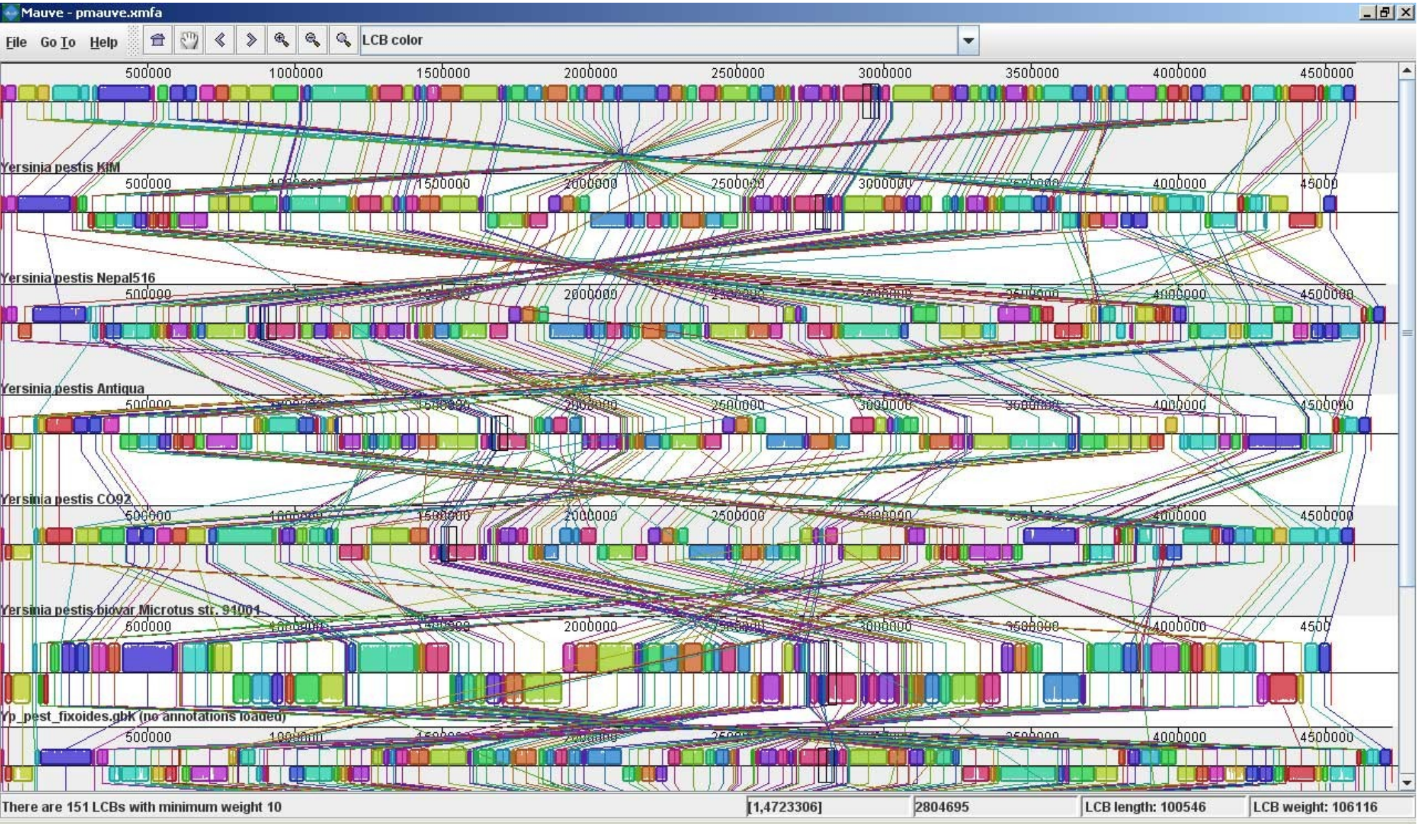
Infernal (RNA Search)

- Infernal is like HMMER
  - Includes use of secondary structure information
  - Uses profile “stochastic context-free grammar”
    - SCFGs vs HMMs
  - “consensus RNA secondary structure profiles”
- Infernal is slow!
- Infernal can be used to search RFAM

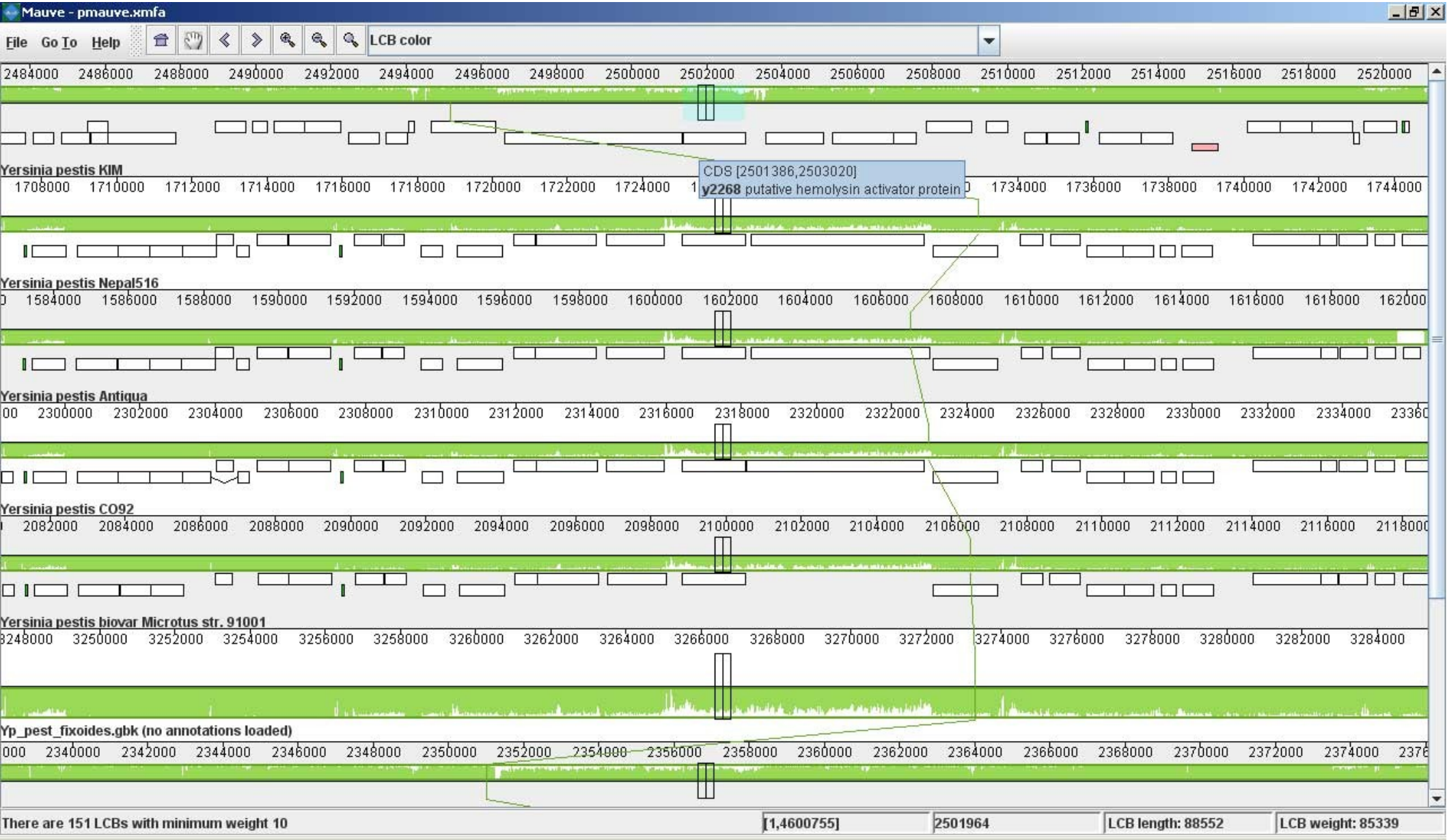
Genome Alignment

- Genome alignment useful for
  - Visualizing genome
    - Rearrangements
    - Insertions/deletions
    - Inversions
  - Annotating genomes
    - Comparing gene annotations across species

Mauve


Mauve - pmauve.xmfa
File
Go To
Help
LCB color
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
Yersinia pestis KIM
Yersinia pestis Nepal516
Yersinia pestis Antiqua
Yersinia pestis CO92
Yersinia pestis biovar Microtus str. 91001
Yp_pest_fixoides.gbk (no annotations loaded)
There are 151 LCBs with minimum weight 10
[1,4723306]
2804695
LCB length: 100546
LCB weight: 106116

Mauve (zoomed in)



Assemblers

- Assemblers job is to make longer sequences from shorter ones.
- Nothing like homology searching
- Must efficiently compare and join billions of sequences
- Soap-Denovo: http://soap.genomics.org.cn/soapdenovo.html
- Amos: http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS
- Many, many, more

Mappers

- These map a read to a reference genome.
- Useful for assembly when a reference genome is already known
  - (think assembly of personal human genomes)
- Identifying SNPs within the same species
- Very Fast!
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- Stampy: http://www.well.ox.ac.uk/project-stampy
- Many Others