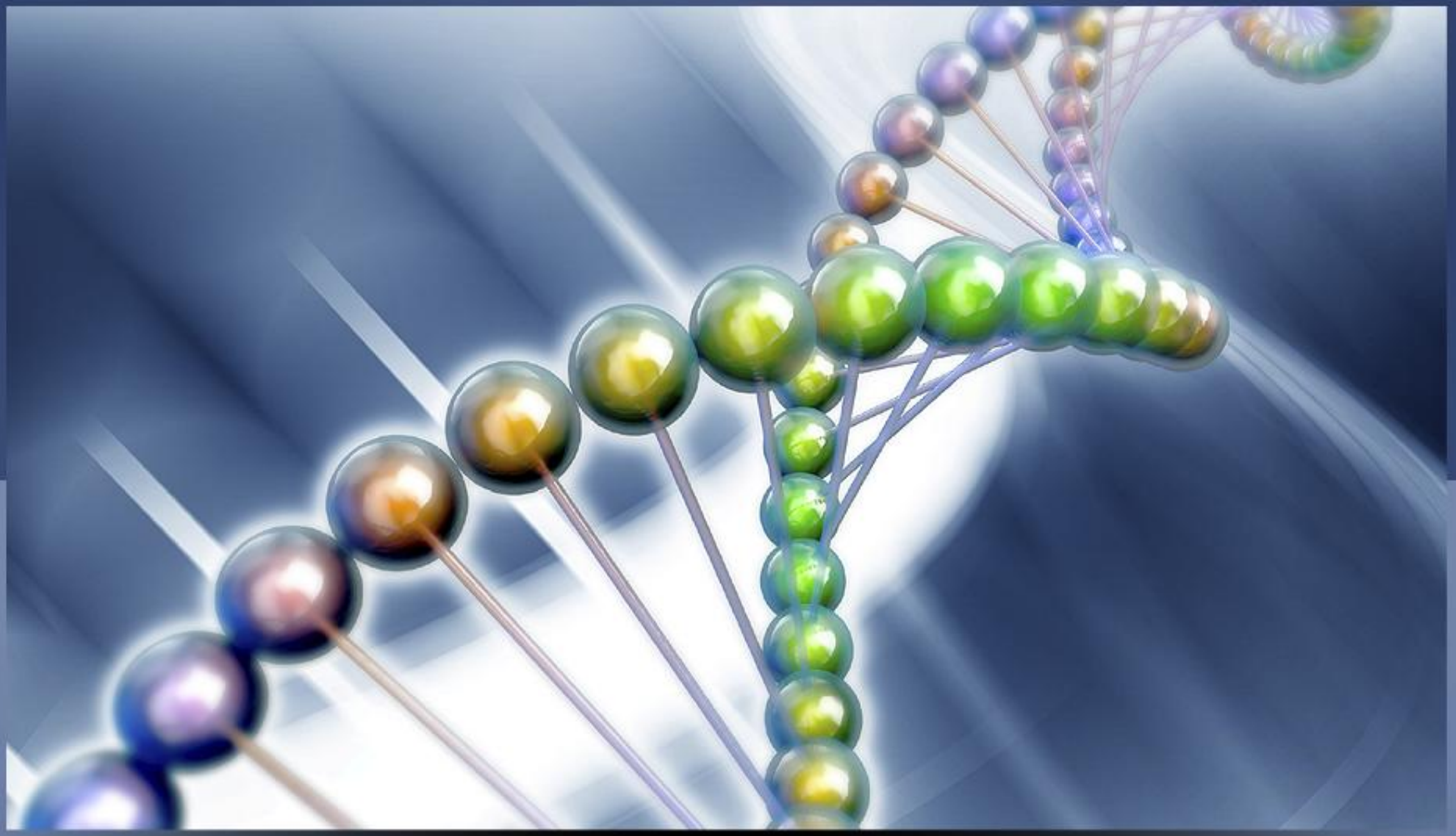
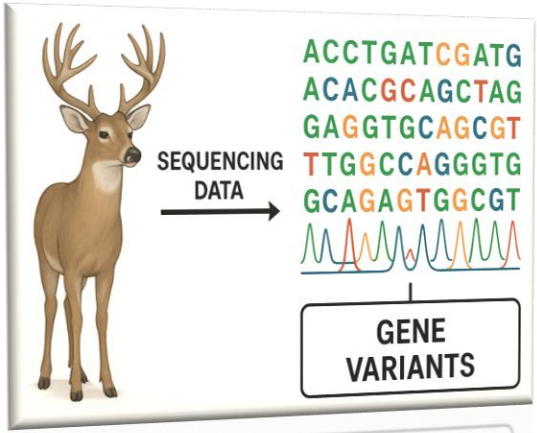
Bioinformatics, FER
Max Henrotin & Lynn Limbach

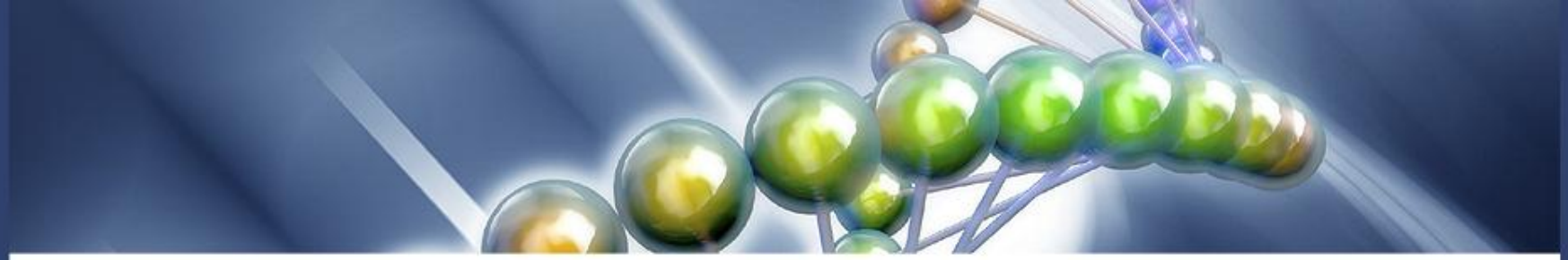


Clustering-Based Identification - Gene Variants in Deer Using Sequencing Data

Introduction



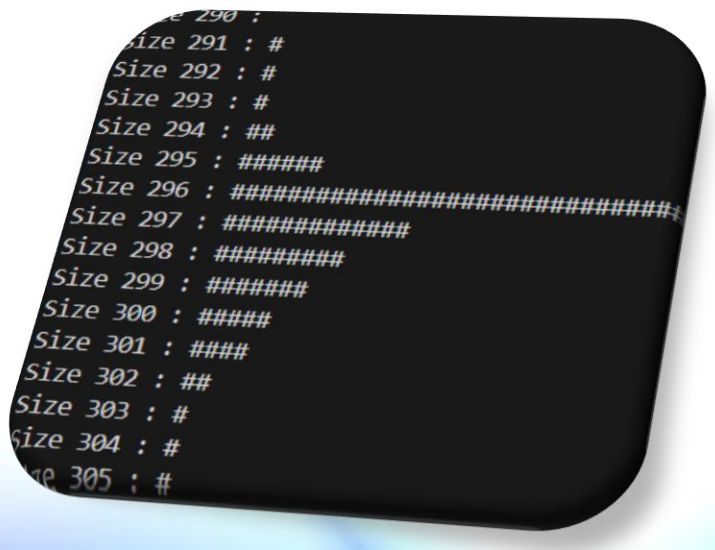
- analysis of gene "Major Histocompatibility Complex"
  - immune system in deer recognizes pathogens
- comparison of appearance in 41 individual deer
  - ~1000 sequence extractions per deer
  - FASTQ format



Our program

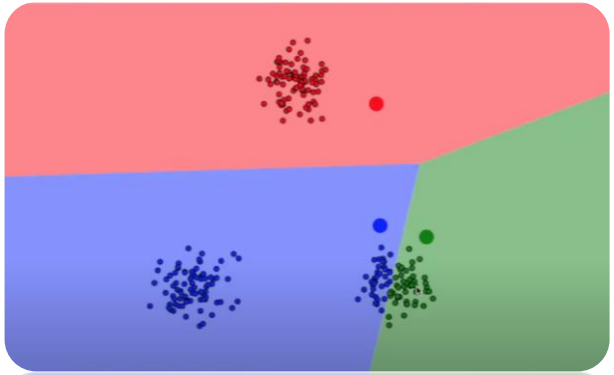
- eliminate corrupted sequences
- correcting measurement errors
- distinguish multiple gene variants
  - k between 2 and 6
  - try to isolate the main variants

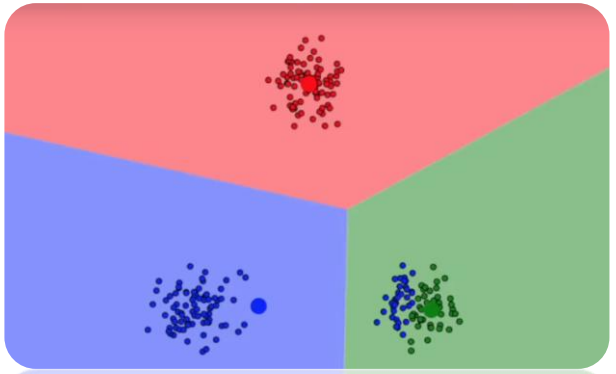
Design & Implementation



- extract FASTQ file into vector of string
- plot histogram to eliminate irrelevant measurements
- implement k-means clustering

K-means clustering

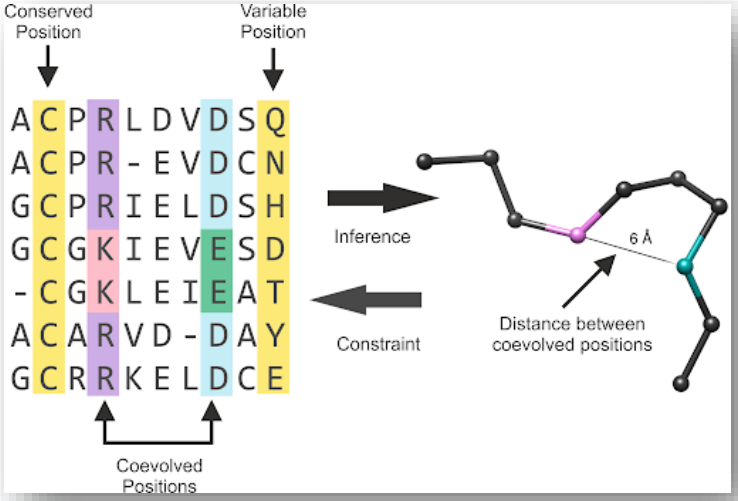




- distance metric
- calculating centroids

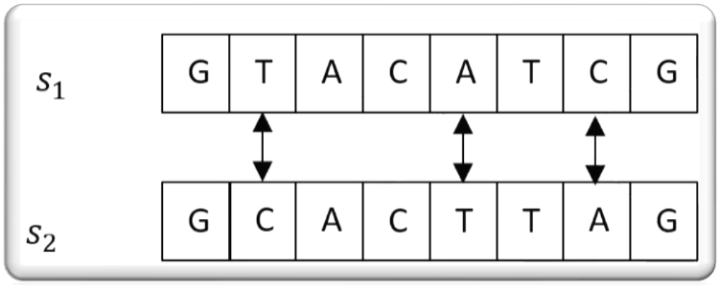
https://www.youtube.com/watch?v=R2e3Ls9H_fc

SPOA for Alignment & Consensus



- MSA alignment
- graph based alignment
  - generate consensus (= centroid)

Distance Metrics



- Levenshtein distance
- Hamming distance

https://www.researchgate.net/publication/365107251/figure/fig1/AS:11431281094739096@1667568473961/Illustration-of-Hamming-Distance-Between-Strings.jpg

Evaluation Strategy

- number of 296 bp sequences per file
  - 200-3500 bp
- amount of clusters
  - 3, 6
- sample file name variations for some deer
  - differences not specified

Observations

- 0 = identical
- digit = difference in bp
- - = too many differences

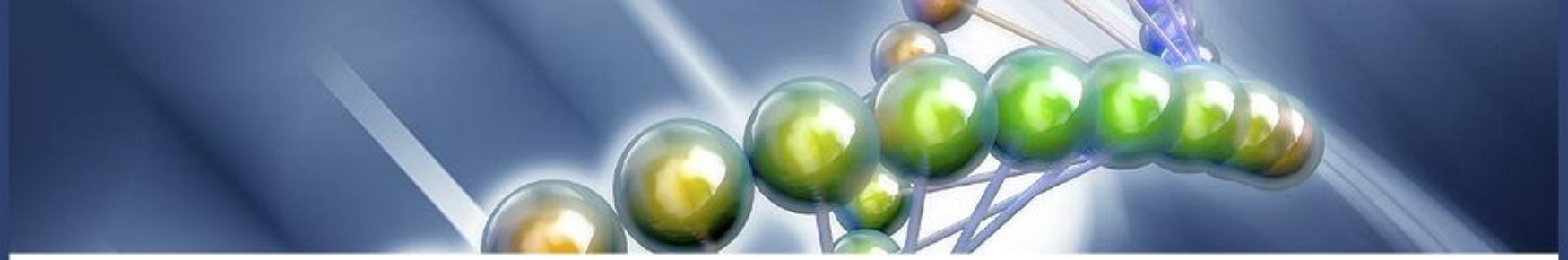
Comparison for: J4_S_CE_IonXpress_021 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 36	-	0	-
cluster no 2 : 276	0	-	-
cluster no 3 : 1	-	1	-
TOTAL TIME : 22.5554 seconds			

Observations

- clustering
  - no difference between amount 3 or 6
  - longer runtime when higher amount
  - 1st cluster found, 2nd with errors, 3rd not found

- matches
  - most with 500+ sequences per file and J30B
  - rare matches in smaller files, none in largest

- similar filenames produce different results



Conclusion

- SP0A-based program with Hamming distance overall produces realistic result
- enough input needed, k=3 sufficient, J30B more common, 3rd cluster not identifiable (confirmed by supervisor), similar filename =/= genetic similarity
- refinements for larger datasets with dynamic k and better prefiltering