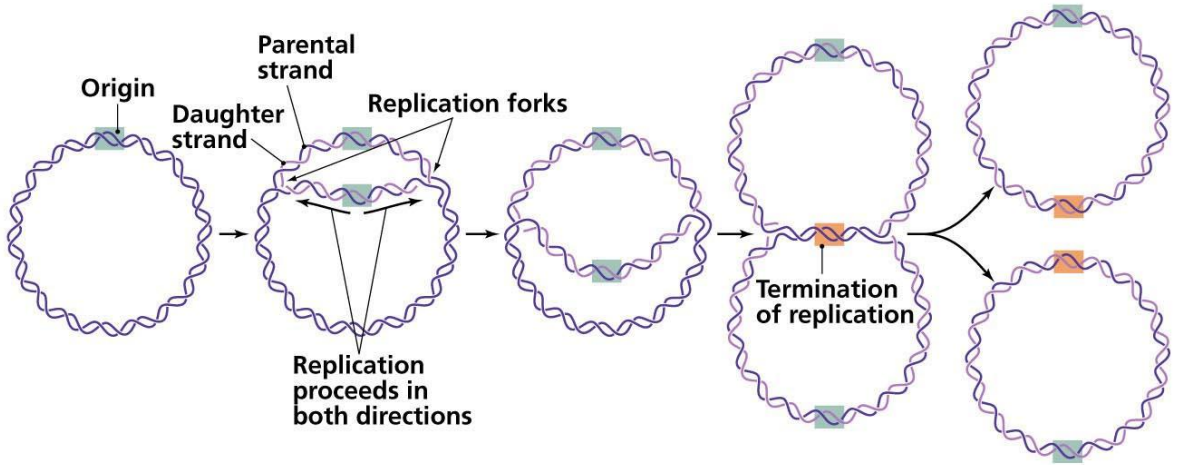
Computational prediction of ORI

ELL796 Project Presentation

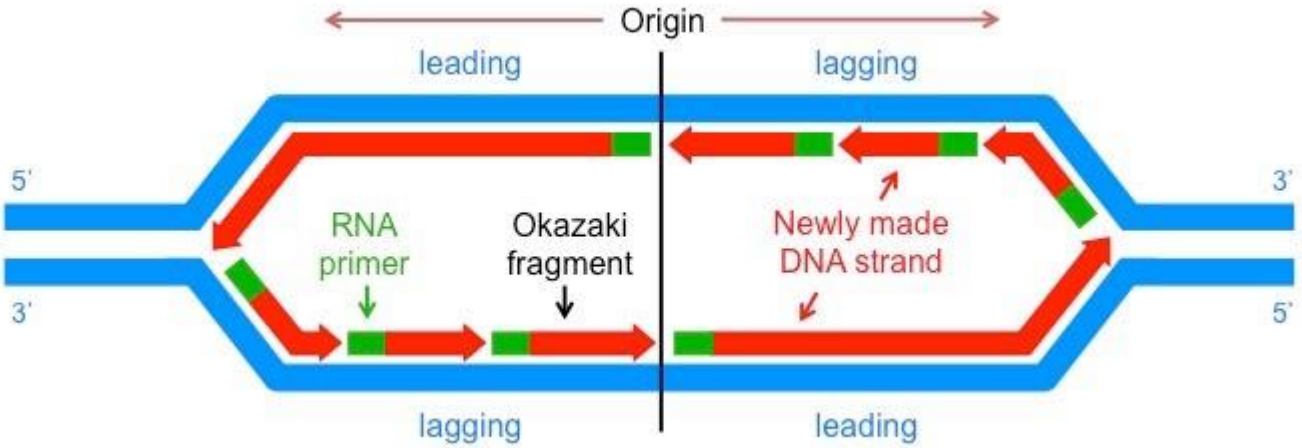
What is ORI?

ORI

1. Point in genome where replication starts
2. Most bacterial genomes are circular
3. Usefiul in studying organisms studying drugs
4. Significantly harder for eukaryotic genomes



Mechanism: Okazaki fragments

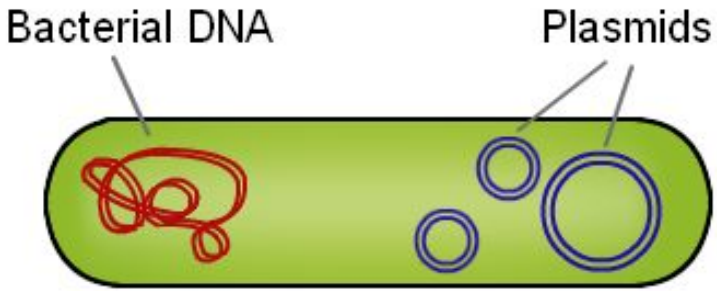


Datasets used

Bacterial genomes:

1. B. subtilis (NC 000964)
2. M. jannaschii (NC 000909)
3. N. tabacum plastid (NC 001879)
4. E. coli (NC 017626)
5. H. Influenzae PittGG, complete genome (CP000672.1)

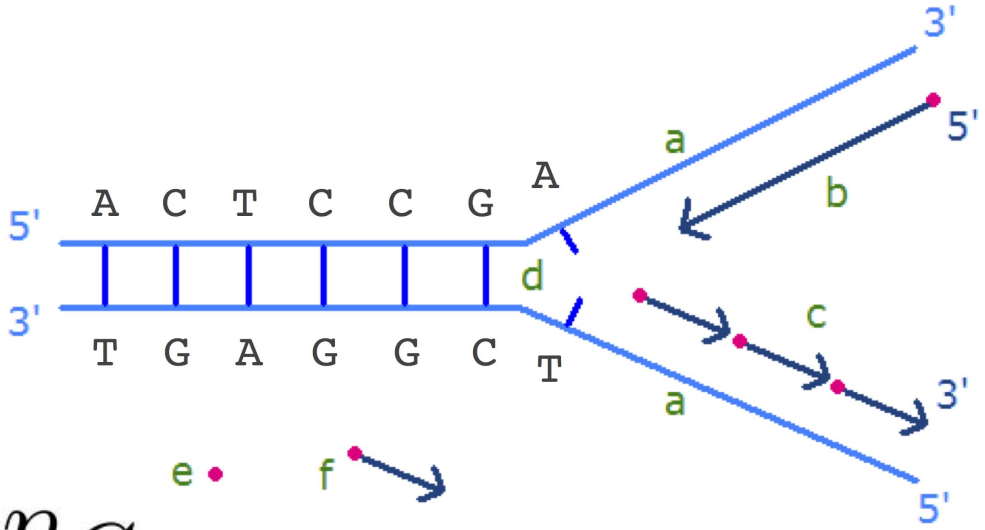
Single circular ‘chromosome’ on which to detect ORI



Cumulative GC skew

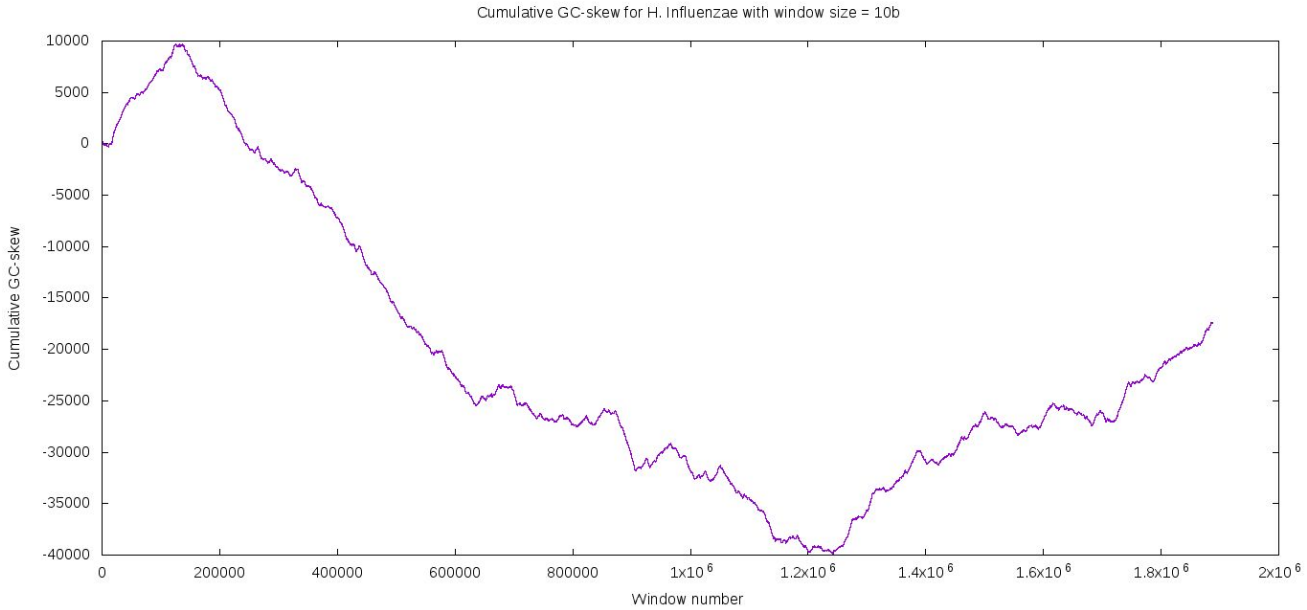
Idea

1. Lagging strand susceptible to mutational pressures
2. C less likely than G
3. GC-skew in a window
4. Add up skews over windows - kind of ‘integration’



$$\frac{n_G - n_C}{n_G + n_C}$$

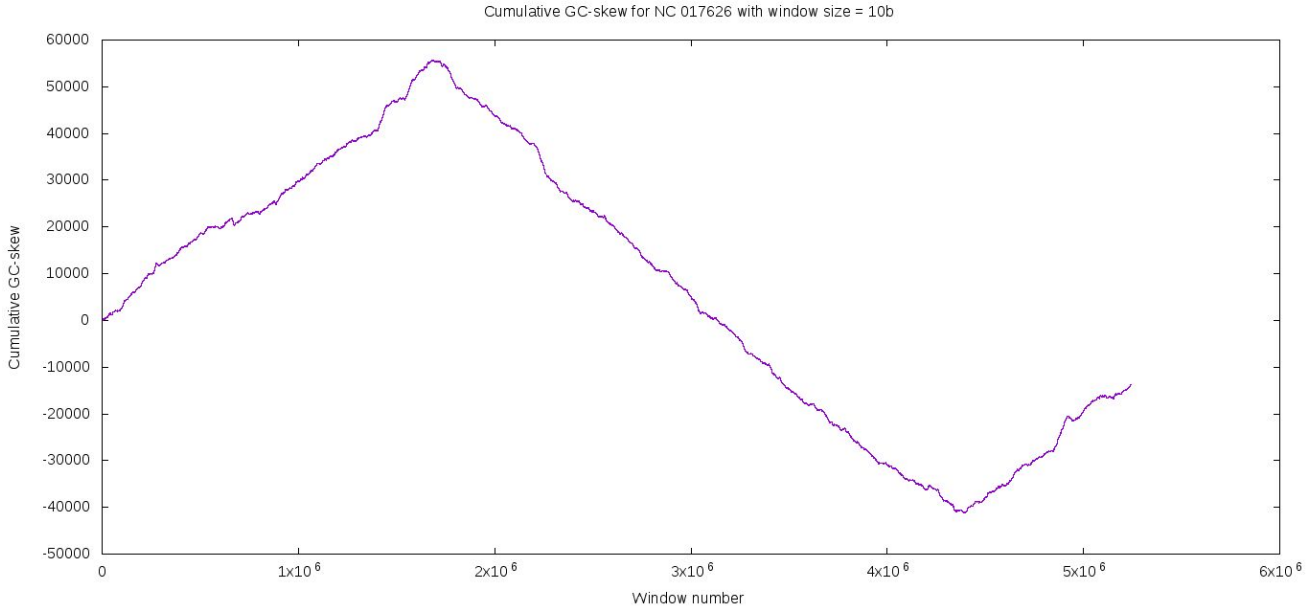
ORI and terminus identification



H. Influenzae

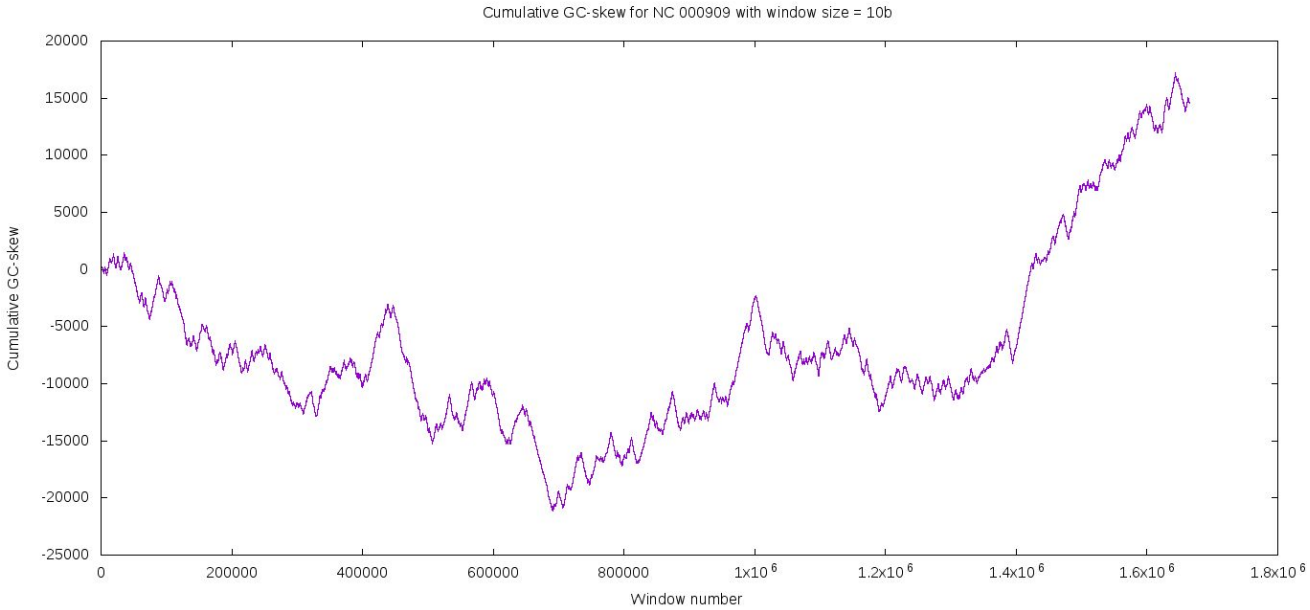
Origin and terminus of replication visible as peak and valley

ORI and terminus identification



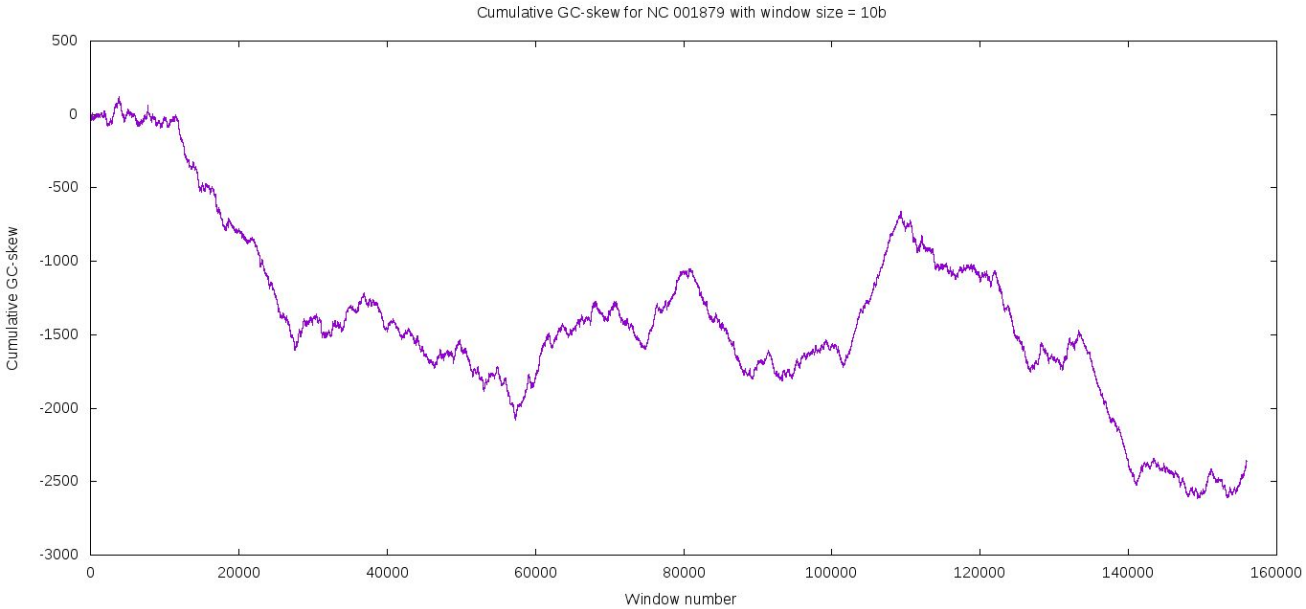
E. Coli

Inconclusive ORI



M. jannaschii

Inconclusive ORI



N. tabacum plastid

Remarks

1. GC skew replicated successfully
2. Performs great on a few bacterial genomes
3. Not so well on others
4. Need for other methods?

Auto-correlation based measures

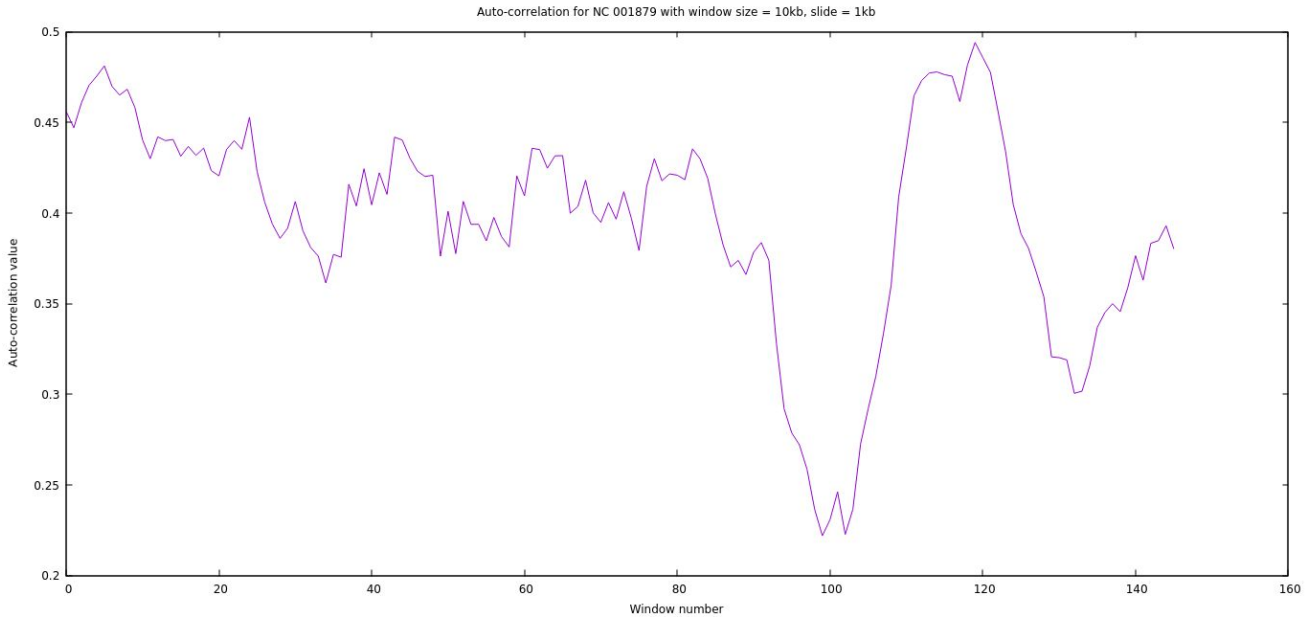
Formulation

1. GC-skew does not always work
2. Idea of correlation to auto-correlation
3. Autocorrelation measure - +1 to one nucleotide, -1 to others
4. iCorr measure - fourth roots of unity to each nucleotide

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} a_j a_{j+k}$$

$$C_G = \frac{1}{N-1} \sum_{k=1}^{N-1} |C(k)|$$

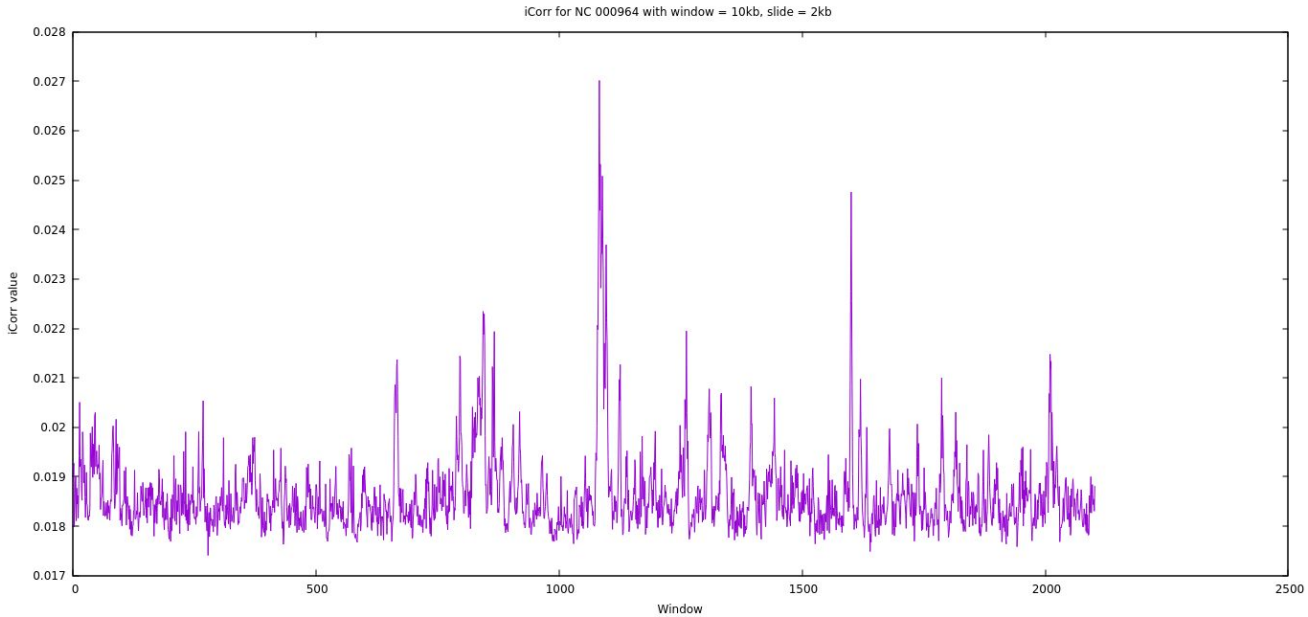
Auto-correlation sudden slope



N. Tabacum Plastid

Origin visible as sharp increase

iCorr peak



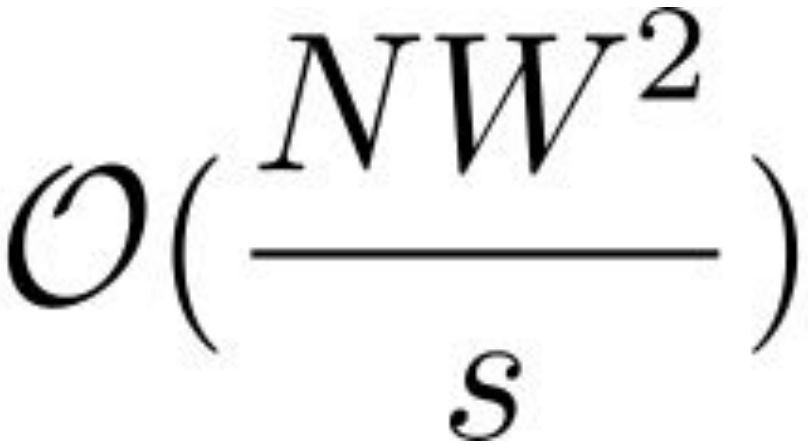
B. subtilis

ORI visible as distinct peak

Computational complexity of implementation

1. N = no. of sequence elements
2. W = window size
3. s = window slide distance

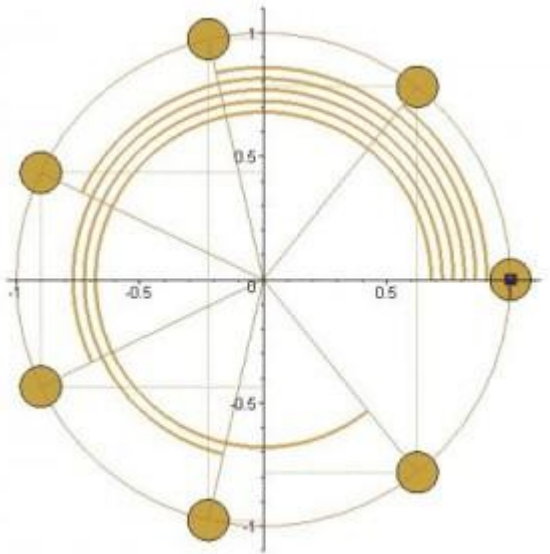
More efficient than naive formula substitution



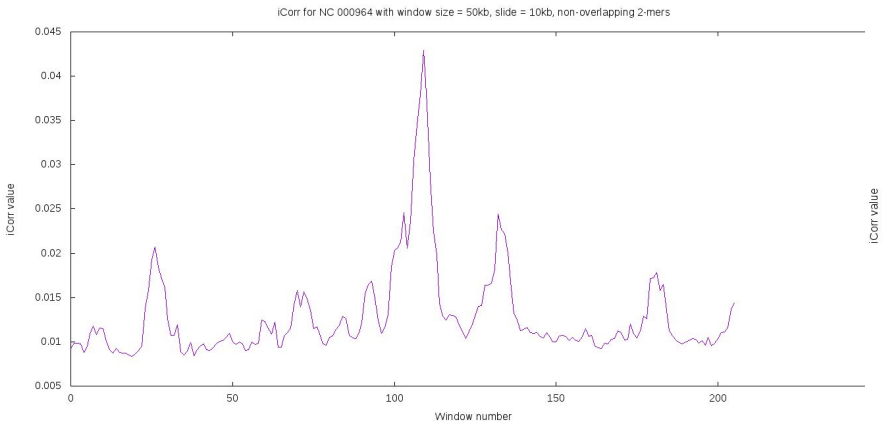
Sequence transformations with k-mers

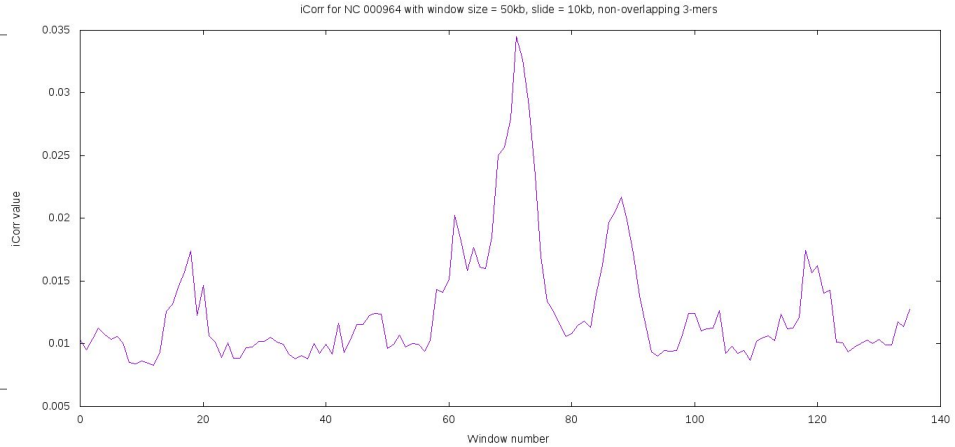
Specifically...

1. 2-mers and 3-mers
2. Map all possible k-mers to roots of unity
3. Apply auto-correlation measures to new sequence



Surprising similarity

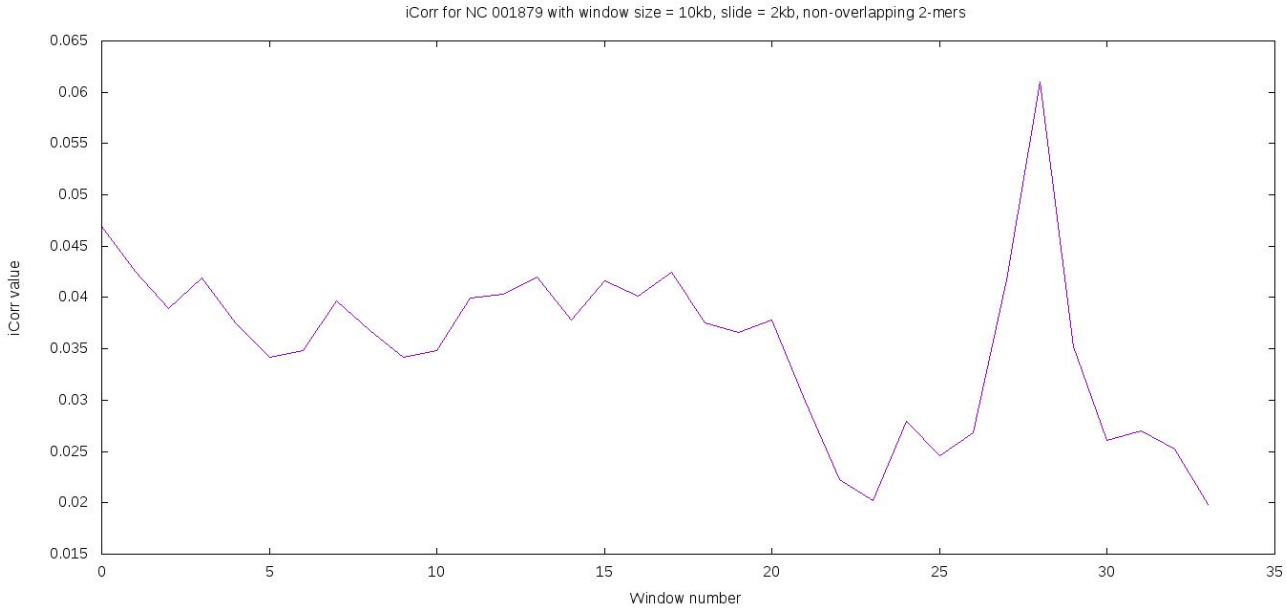




N. Tabacum Plastid

2-mer and 3-mer plot very similar

Similar to auto-correlation

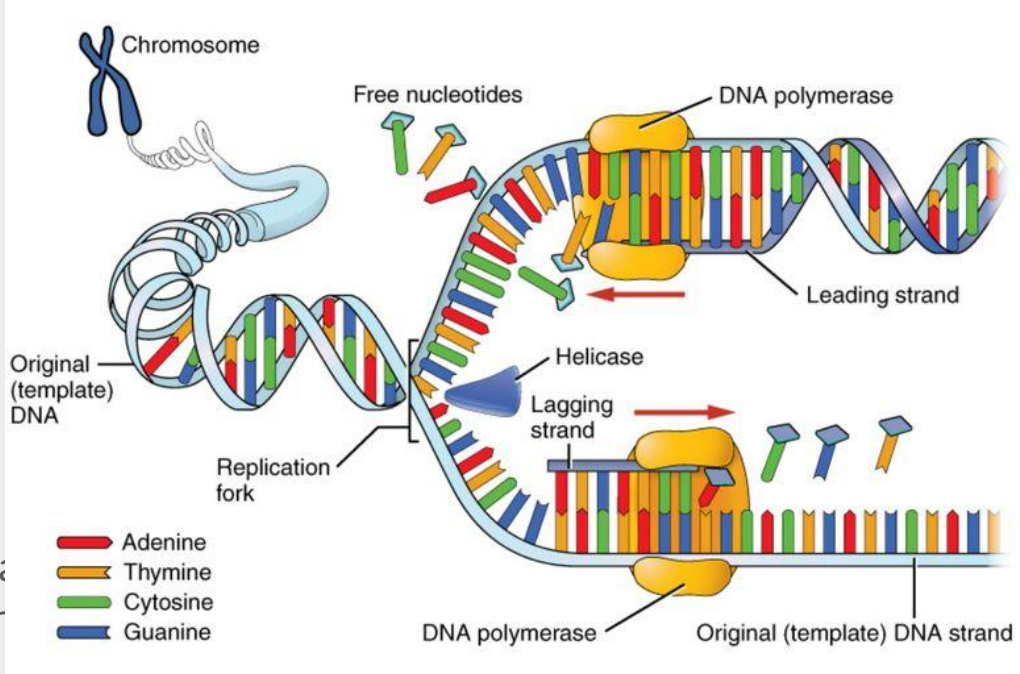


N. Tabacum plastid

Implies same nucleotides likely to be followed by same nucleotide?

Conclusion

1. Studied ORI prediction techniques
2. Successfully replicated existing methods
3. Plots for transformed sequences often similar between k-mers
4. ‘Anomalous’ cosecutive pea
worth further investigation



Thank you

Abhishek Pathak
2015CS10424

cs1150424@iitd.ac.in