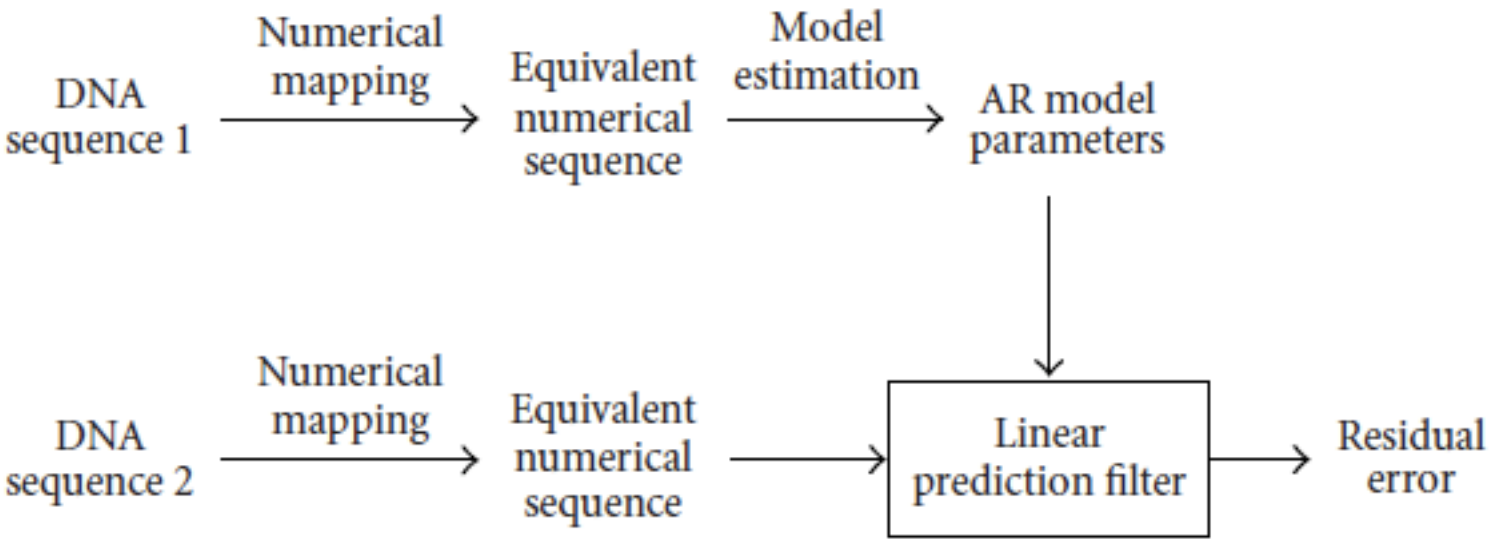
AR Modeling

ECE-S690



Important

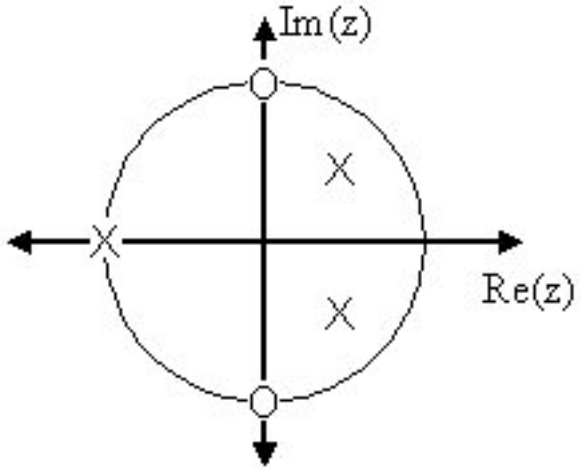
- Literature Review and Project Proposals – NEXT WEEK

- 30 minute presentations (20 minutes on background, problems, methods and 10 minutes on proposal)

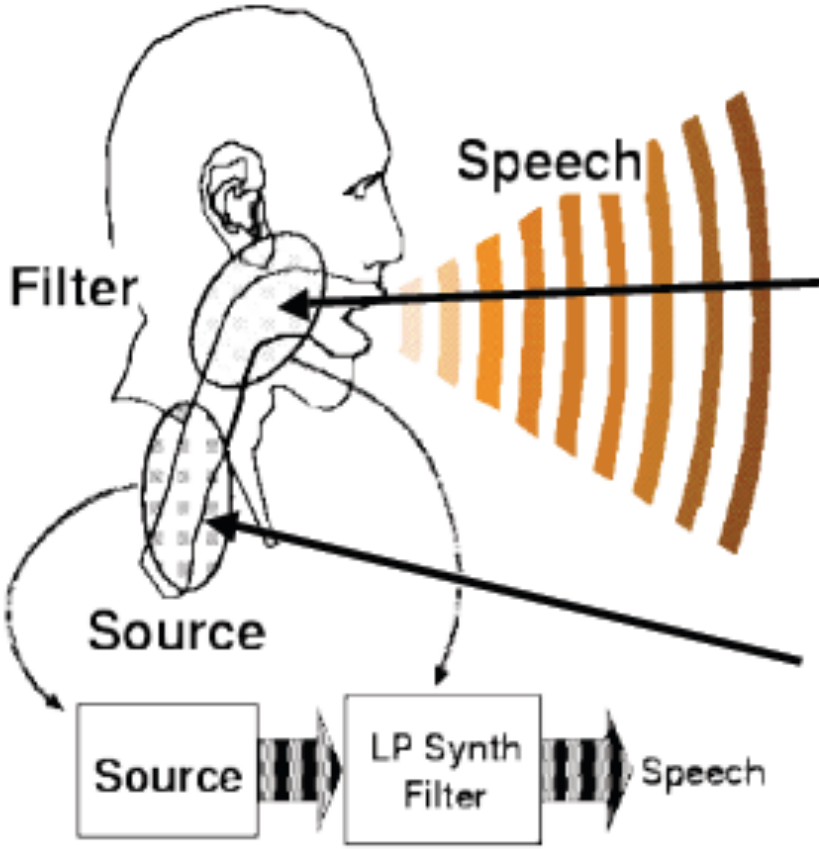
ARMA Modeling

- Y(z)=H(z)X(z)
- Y(z)/X(z)=H(z) (Input/Output)
- H(z)=B(z)/A(z)

- A(z) models poles
- B(z) models zeros



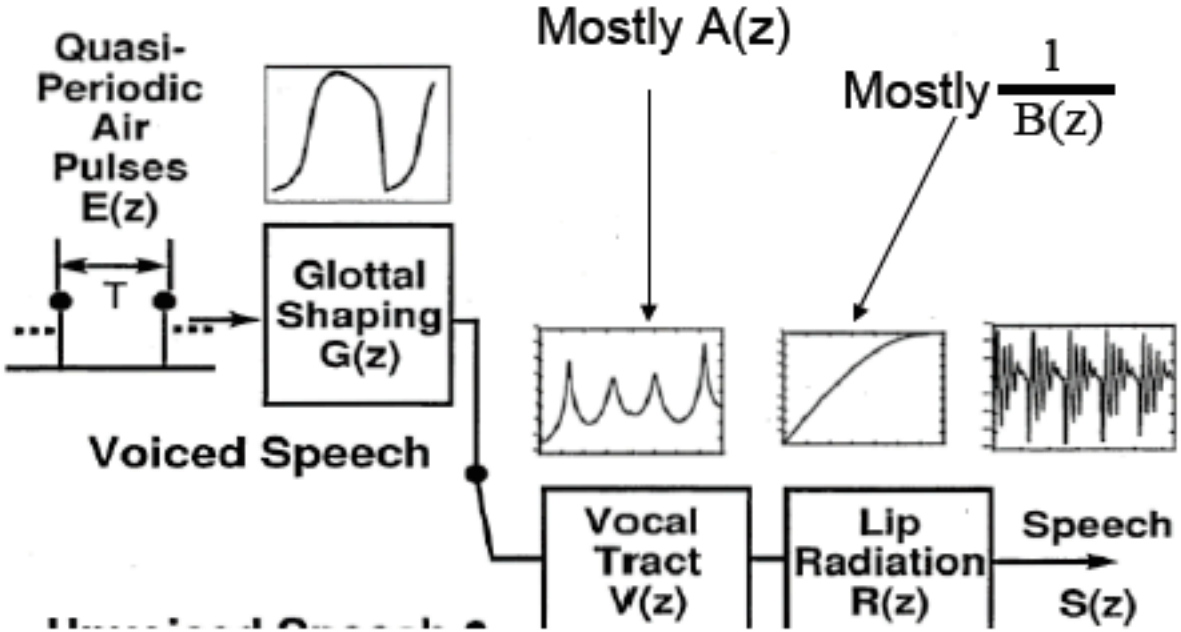
Example: Speech Processing



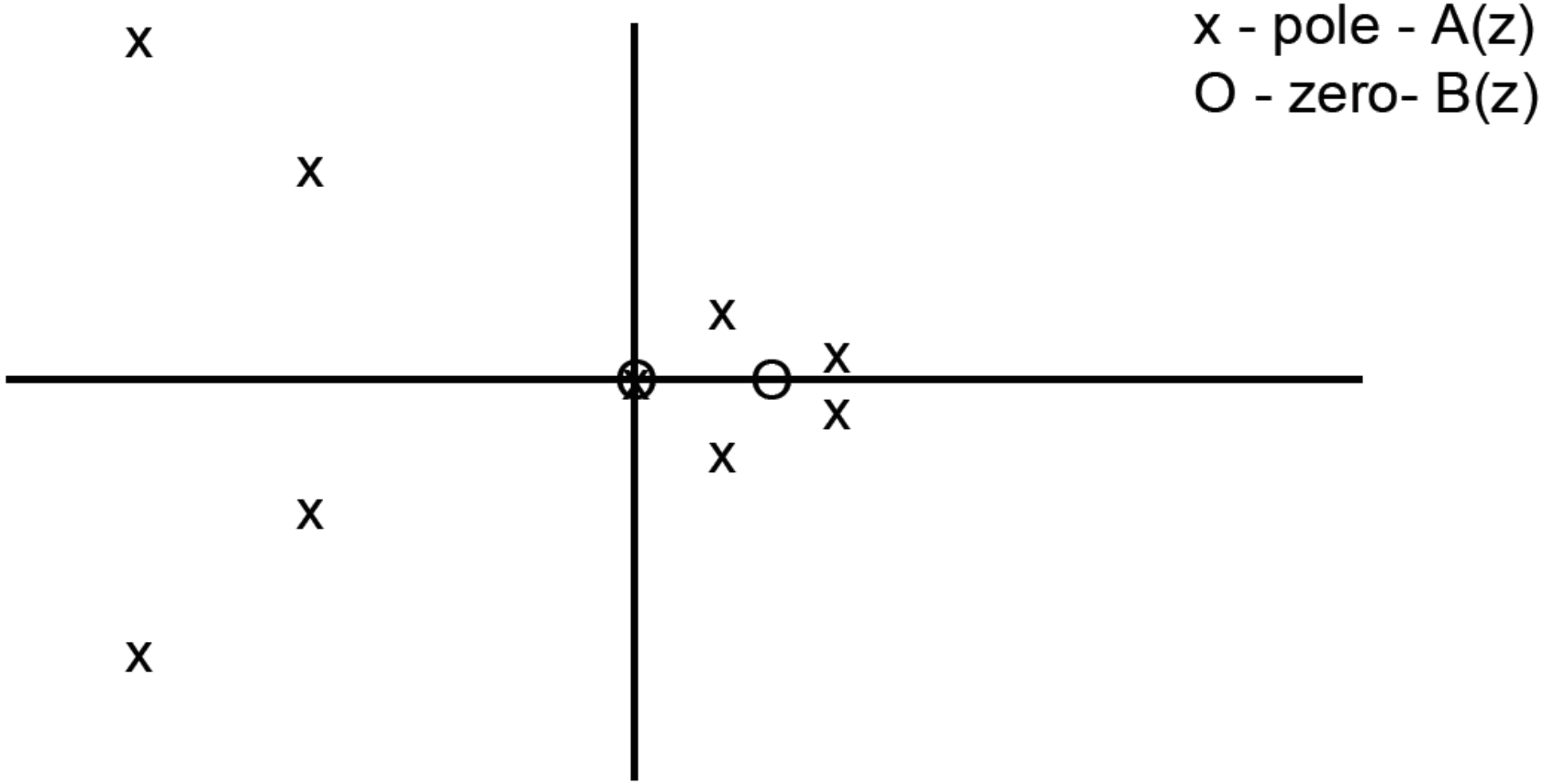
Vocal Chords: A(z) filter -- poles

Excitation: Glottal Pulses

Speech little more complex…but



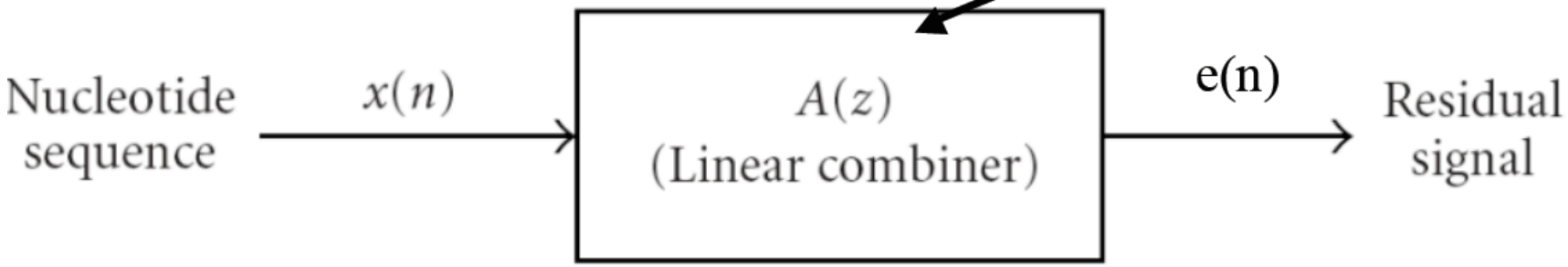
Pole-Zero Plots



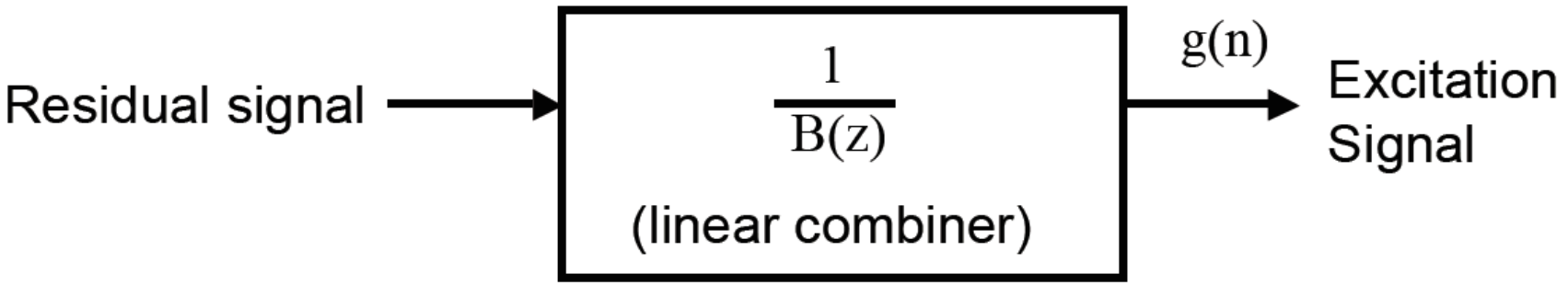
ARMA Modeling

- AR = autoregressive

a.k.a. Linear Prediction Filter



- MA = moving average



ARMA Modeling

- A(z) can be approximated by a coefficients
- B(z) can be approximated by b coefficients

Time/Frequency Domain

$$y(n) = \sum_{m=1}^{N} a_m y(n-m) + \sum_{m=0}^{M} b_m x(n-m)$$

$$\frac{Y(e^{j\omega})}{X(e^{j\omega})} = H(e^{j\omega}) = \frac{B(e^{j\omega})}{A(e^{j\omega})} = \frac{\sum_{m=0}^{M} b_m e^{-j\omega m}}{1 - \sum_{m=1}^{N} a_m e^{-j\omega m}}$$

Autocorrelation review

$$r_{xx}(m) = E\left[x(n+m)x(n)\right]$$

Stationary, Ergodic

$$= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n+m)x(n),$$

Classical Estimator

$$\hat{r}_b(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n+|m|)x(n)$$

Symmetry

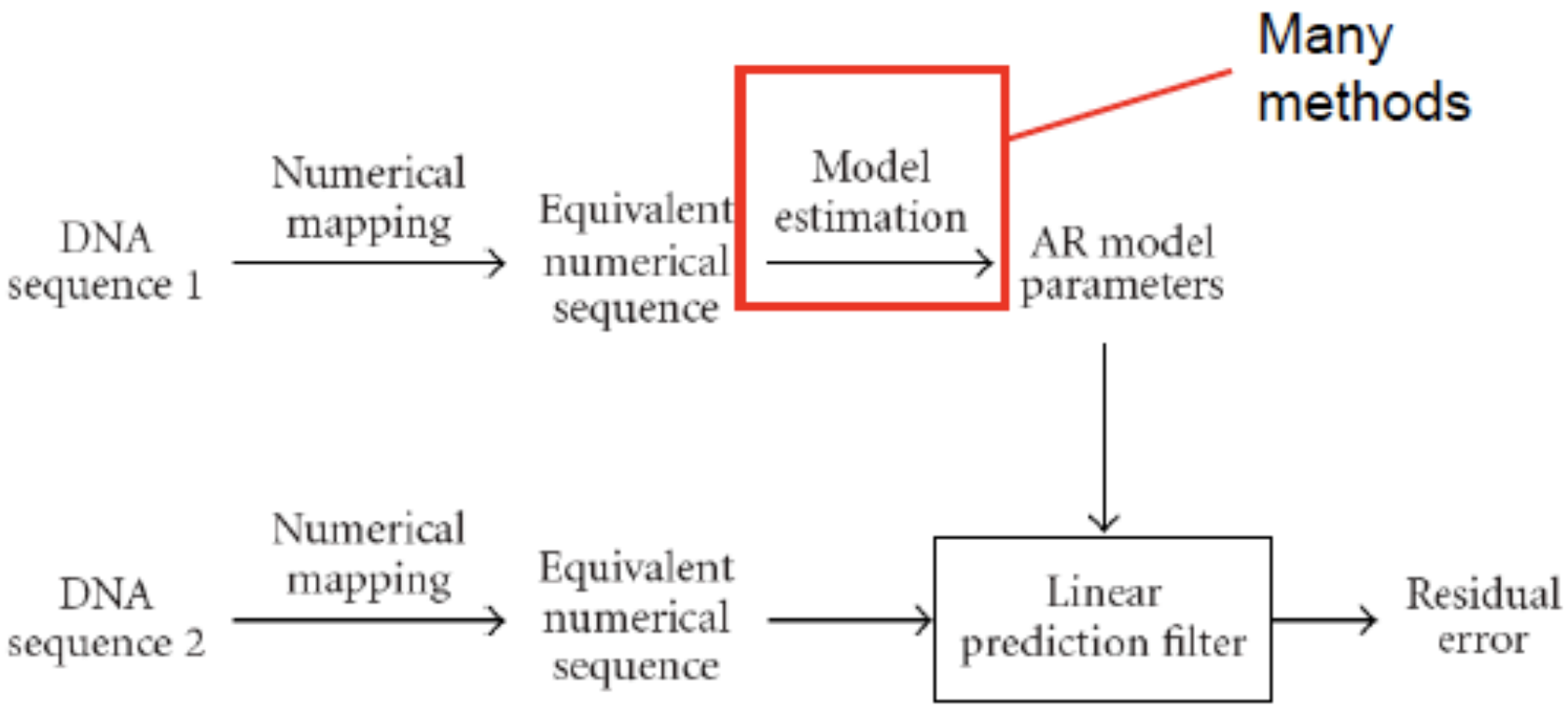
Power Spectrum relations

$$r_{xx}(m) \quad \xleftrightarrow{\text{Fourier Transform}} \quad P(e^{j\omega})$$

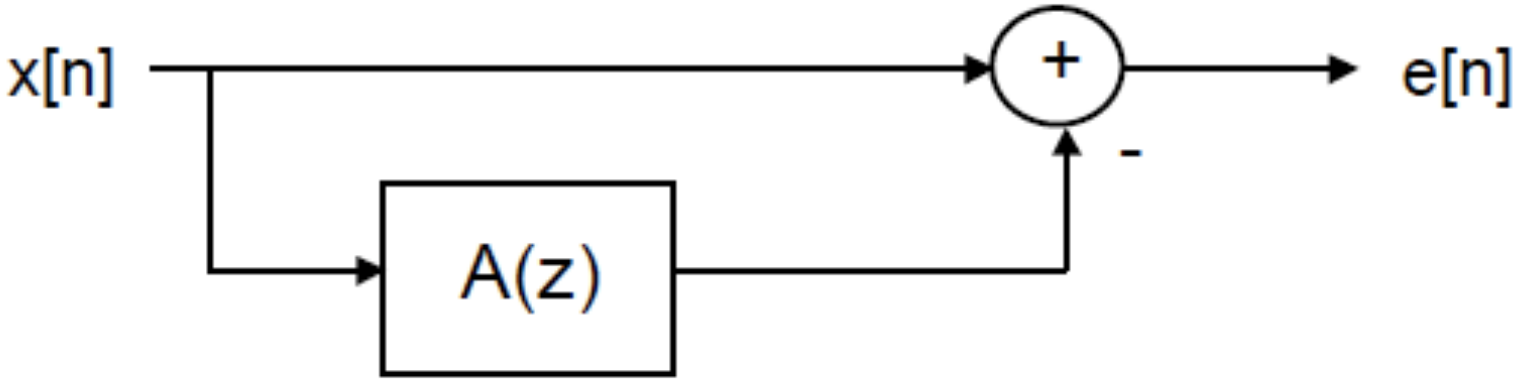
Transform of Autocorrelation is Power Spectrum

$$P(e^{j\omega}) = \left| \frac{B(e^{j\omega})}{A(e^{j\omega})} \right|^2$$

DNA AR modeling



Linear Prediction Analysis



Model Poles only -- Inverse Filtering

Yule-Walker: Popular Way to get ***a*** coefficients

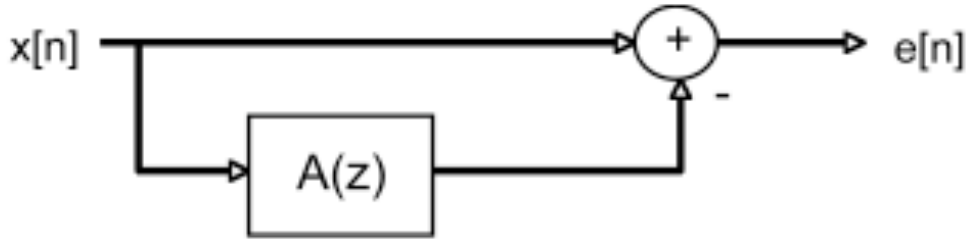
$$\begin{pmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \cdots & R_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = -\begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{pmatrix}$$

$$\begin{pmatrix} \hat{R}_0 & \hat{R}_1 & \cdots & \hat{R}_{p-1} \\ \hat{R}_1 & \hat{R}_0 & \cdots & \hat{R}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}_{p-1} & \hat{R}_{p-2} & \cdots & \hat{R}_0 \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix} = -\begin{pmatrix} \hat{R}_1 \\ \hat{R}_2 \\ \vdots \\ \hat{R}_p \end{pmatrix}$$

$$\hat{R}_\tau \equiv \frac{1}{N} \sum_{t=\tau+1}^{N} y_t y_{t-\tau}$$

AKA: Levinson-Durbin

Minimize Error



$$\frac{X(z)}{E(z)} = \frac{1}{1 - A(z)} = \frac{1}{1 - \sum_{m=1}^{N} a_m z^{-m}}$$

$$e[n] = x[n] - \sum_{m=1}^{N} a_m x[n-m]$$

Derivation of Mean Square Error (MSE)

$$
\begin{aligned}
E &= \sum_{n=0}^{N-1} e_n^2 \\
&= \sum_{n=0}^{N-1} \left(s_n - \sum_{i=1}^{p} a_i s_{n-i} \right)^2 \\
&= \sum_{n=0}^{N-1} \left(s_n^2 - 2\sum_{i=1}^{p} a_i s_n s_{n-i} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j s_{n-i} s_{n-j} \right) \\
&= \sum_{n=0}^{N-1} s_n^2 - 2\sum_{i=1}^{p} a_i \sum_{n=0}^{N-1} s_n s_{n-i} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sum_{n=0}^{N-1} s_{n-i} s_{n-j} \\
&= \sum_{n=0}^{p} \phi_{00} - 2\sum_{i=1}^{p} a_i \phi_{0i} \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \phi_{ij} \\
&= \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix} \underbrace{\begin{bmatrix} \phi_{00} & \phi_{01} & \phi_{02} & \cdots & \phi_{0p} \\ \phi_{10} & \phi_{11} & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{20} & \phi_{21} & \phi_{22} & \cdots & \phi_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \phi_{p0} & \phi_{p1} & \phi_{p2} & \cdots & \phi_{pp} \end{bmatrix}}_{R} \begin{bmatrix} -1 \\ a_1 \\ a_2 \\ \cdots \\ a_p \end{bmatrix}
\end{aligned}
$$

R (autocorrelation of Frame 1)

Energy/MSE = $a^T R a$

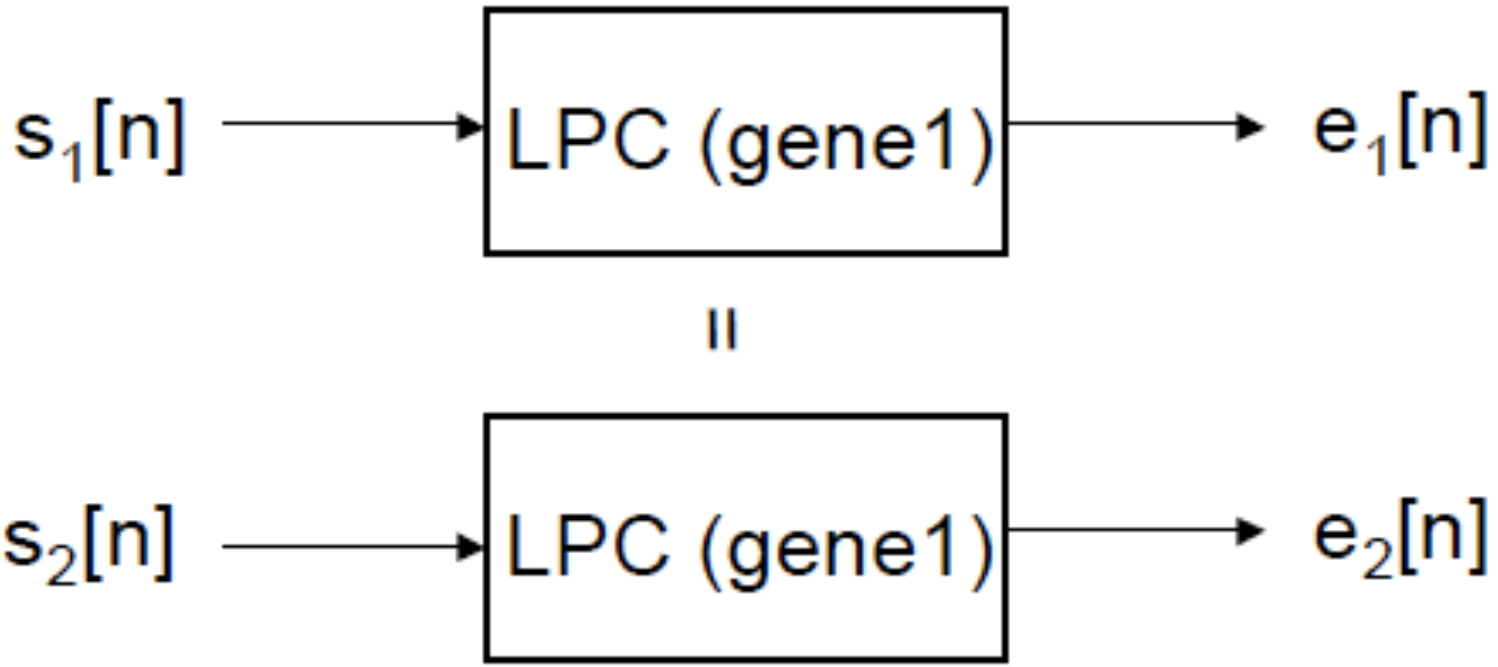
$E_a = a^T R a$

$E_b = b^T R b$

LP and AR modeling Matlab Tutorial

- http://www.mathworks.com/products/signal/demos.html?file=/products/demos/shipping/signal/lpcardemo.html#10

Analysis 1



Disadvantage: if compare e's, two different signals may yield the same

Less Variance == better fit

Rosen, Gensips 2007

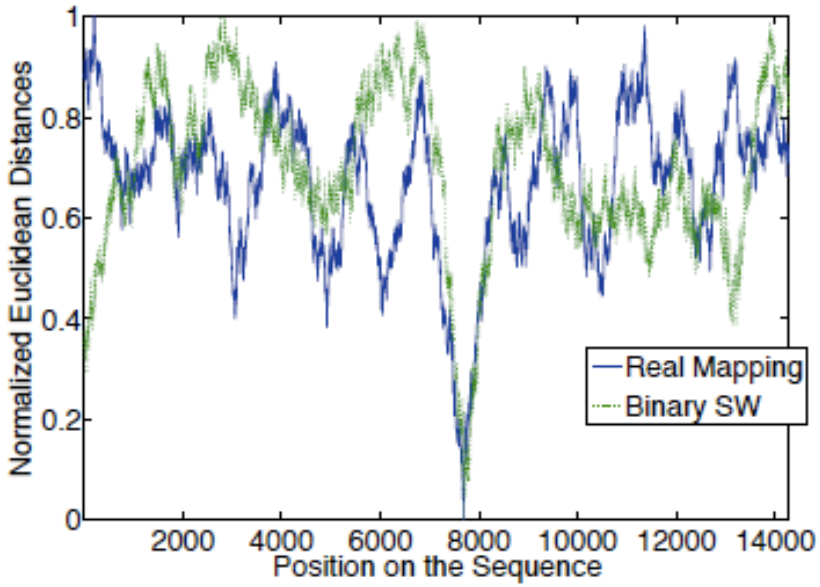
Performance of DNA Representations

Real Representation:

$$A = 1.5, C = 0.5, G = -0.5, T = -1.5$$

Binary (A+T) Rule:

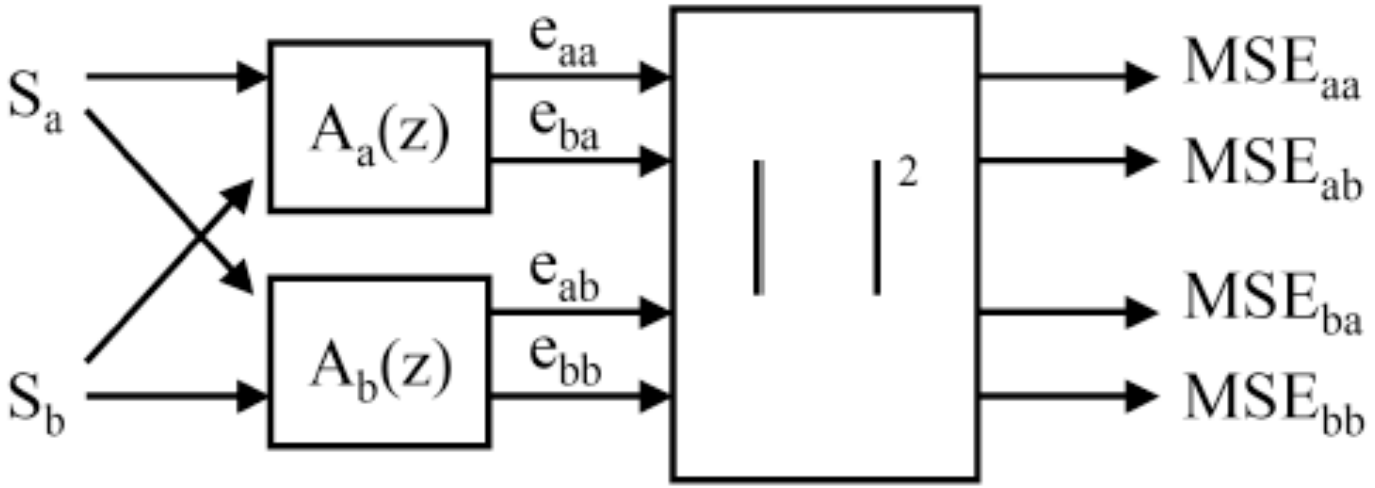
$$A = 1, C = 0, G = 0, T = 1$$



The Real vs. Binary A+T mapping for the Euclidean distance between the exon's and each sequence window's AR coefficients; the sequence window length is the length of the exon. Shown is a portion of S. Cerevisiae chromosome XIV. The exon is located at $7682 \rightarrow 8404$ within this portion and is modeled with an AR order of $p = 14$.

Distance Measures

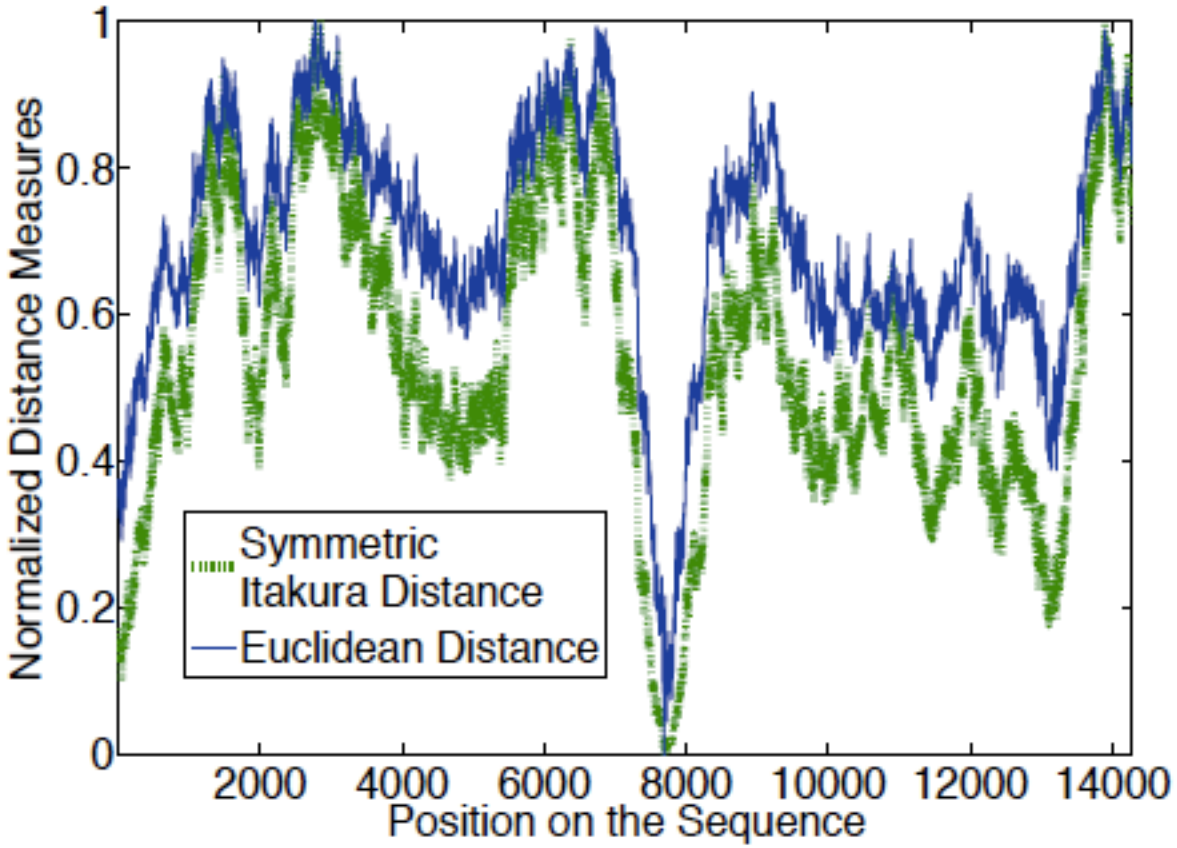
Itakura Distance:



$$d_i(S_a, S_b) = log_{10} \frac{A_b^T R_a A_b}{A_a^T R_a A_a} = log_{10} \frac{MSE_{ab}}{MSE_{aa}}$$

Euclidean Distance:

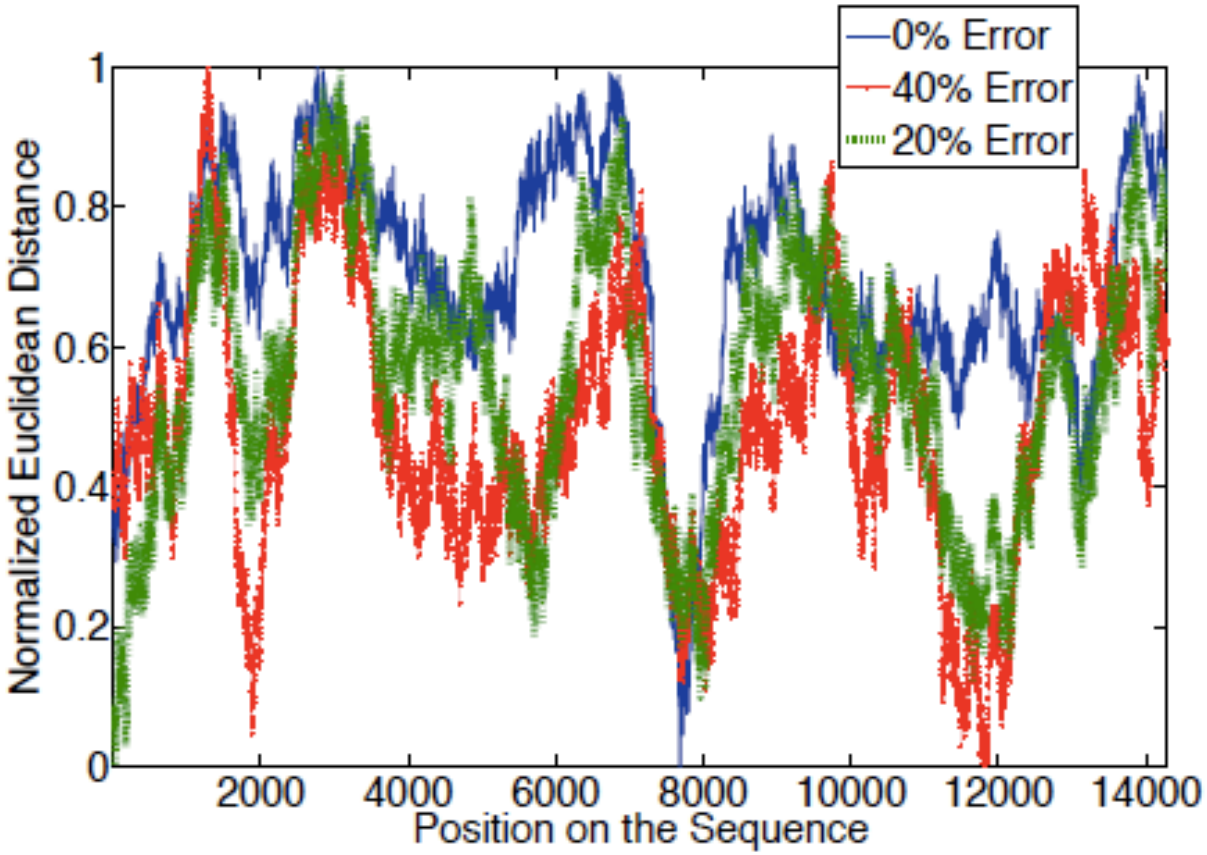
$$d_e(S_a, S_b) = \sqrt{\sum_{i=1}^{p} (a_a(i) - a_b(i))^2}$$



The Euclidean vs. the Itakura distance with the Binary A+T mapping, using the same S. Cerevisiae sequence and same model order of $p = 14$.

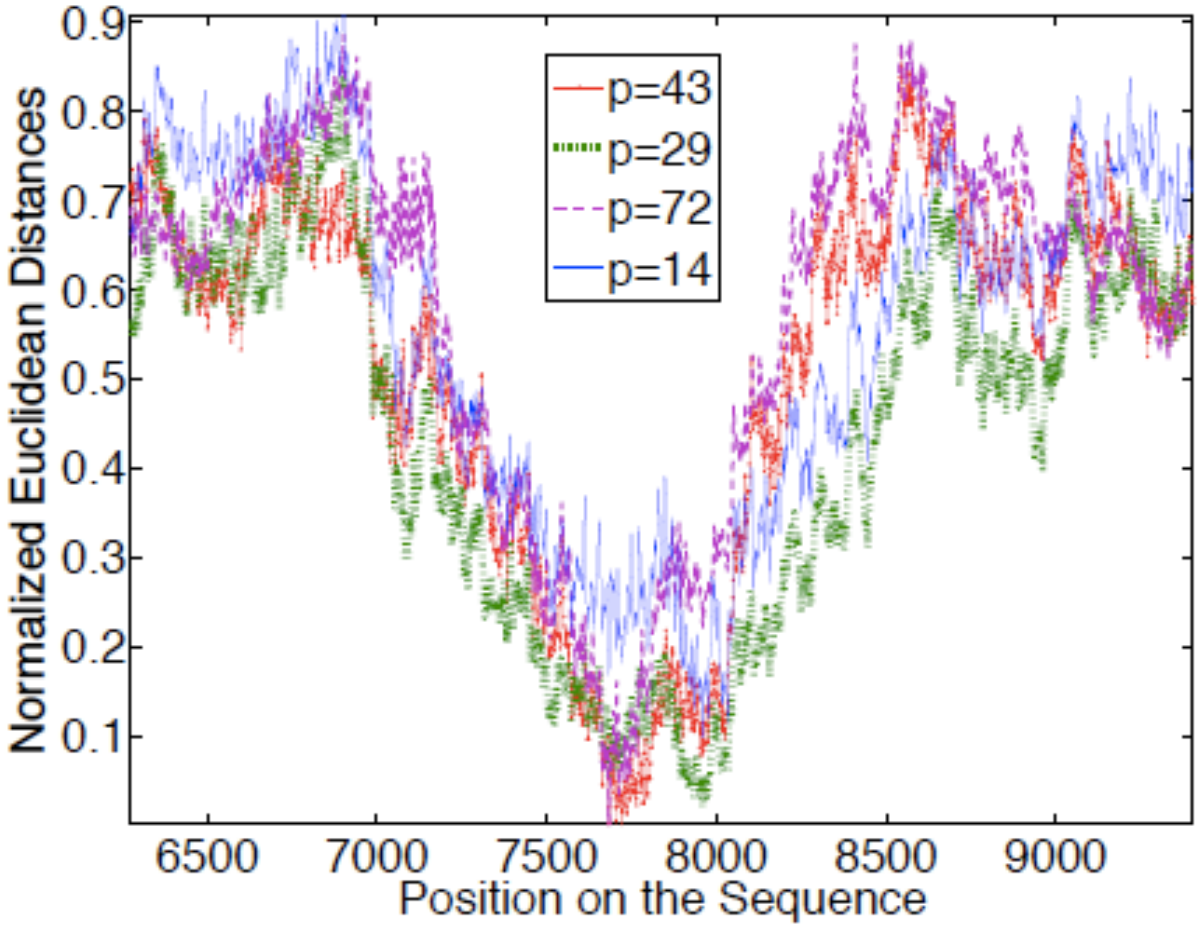
Performance on Perturbed Sequences

Effect of increasing error:



AR Euclidean distance performance vs. percentage mutation rate for model order $p = 14$ on the S. Cerevisiae sequence. A Binary A+T mapping is used.

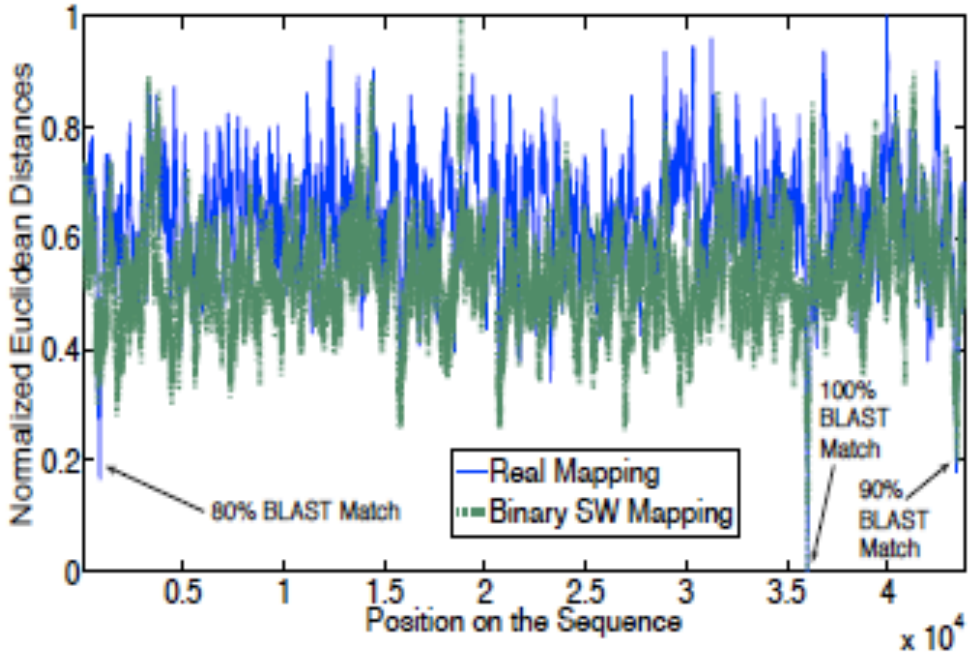
Increasing model order becomes more robust to error:



AR Euclidean distance performance vs. model order for a 20% mutation rate on the S. Cerevisiae sequence. A Binary A+T mapping is used.

Real Sequences

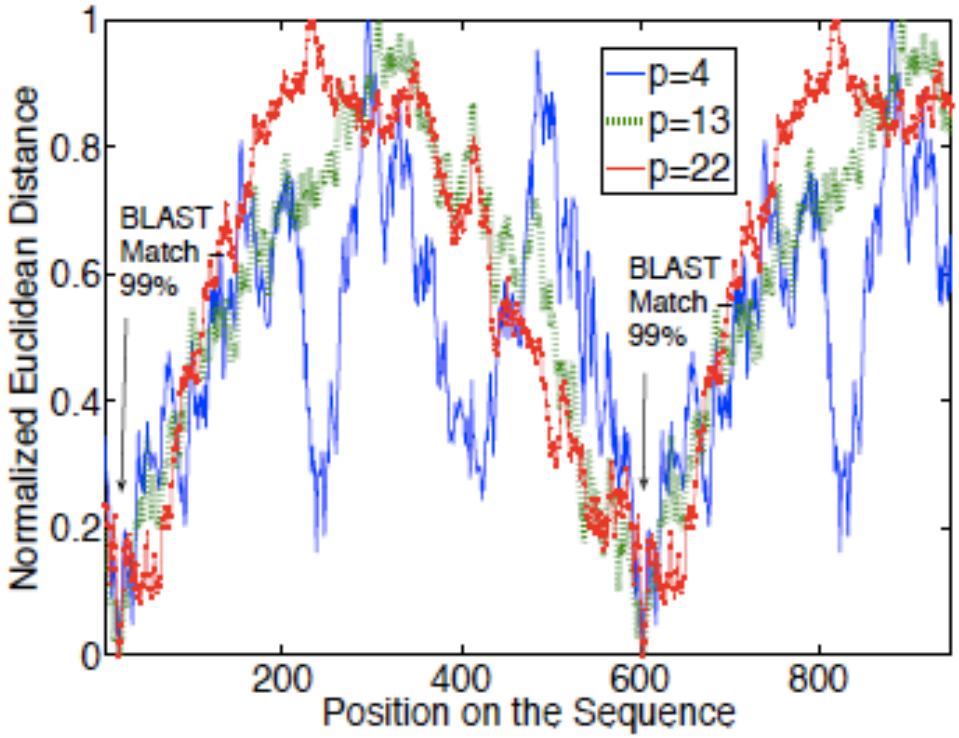
Human Hemoglobin Delta (HHD) exon:



Performance of Euclidean distance for $p = 72$ AR model order vs. mapping for matching a Human Hemoglobin Delta exon (Genbank Accession EF051731, nucleotides $290 \rightarrow 512$) to a Human Beta Globin Region on Chromosome 11 (Genbank Accession U01317.1, nucleotides $19000 \rightarrow 63000$). The real mapping is used.

Real Sequences

HHD vs. Human mRNA:



Performance of Euclidean distance AR model order for matching a Human Hemoglobin Delta exon (Genbank Accession EF051731, nucleotides $290 \rightarrow 512$) to a Human clone Affy08244A08 (mRNA)(Genbank Accession DQ655982.1). The real mapping gave the best match distinction.

Conclusions

- The Numerical Mapping has no effect on the AR similarity measure.
- The Euclidean distance presents greater divergence between the matching and non-matching regions, as opposed to the Itakura distance.
- AR method robust to high error-rates.
- Increasing Model Order improves accuracy, although at high computational cost.
- Method works well on matching real exon regions (known 3-base periodic).
- Trade-off: method is computationally intensive.
- Need: Model order selection for accuracy.

Chakravarthy Paper

Analysis 2

- A(z) coefficients -- Feature vector

$$\boldsymbol{a} = [1\ a_1\ a_2\ a_3\ \ldots\ a_N]$$

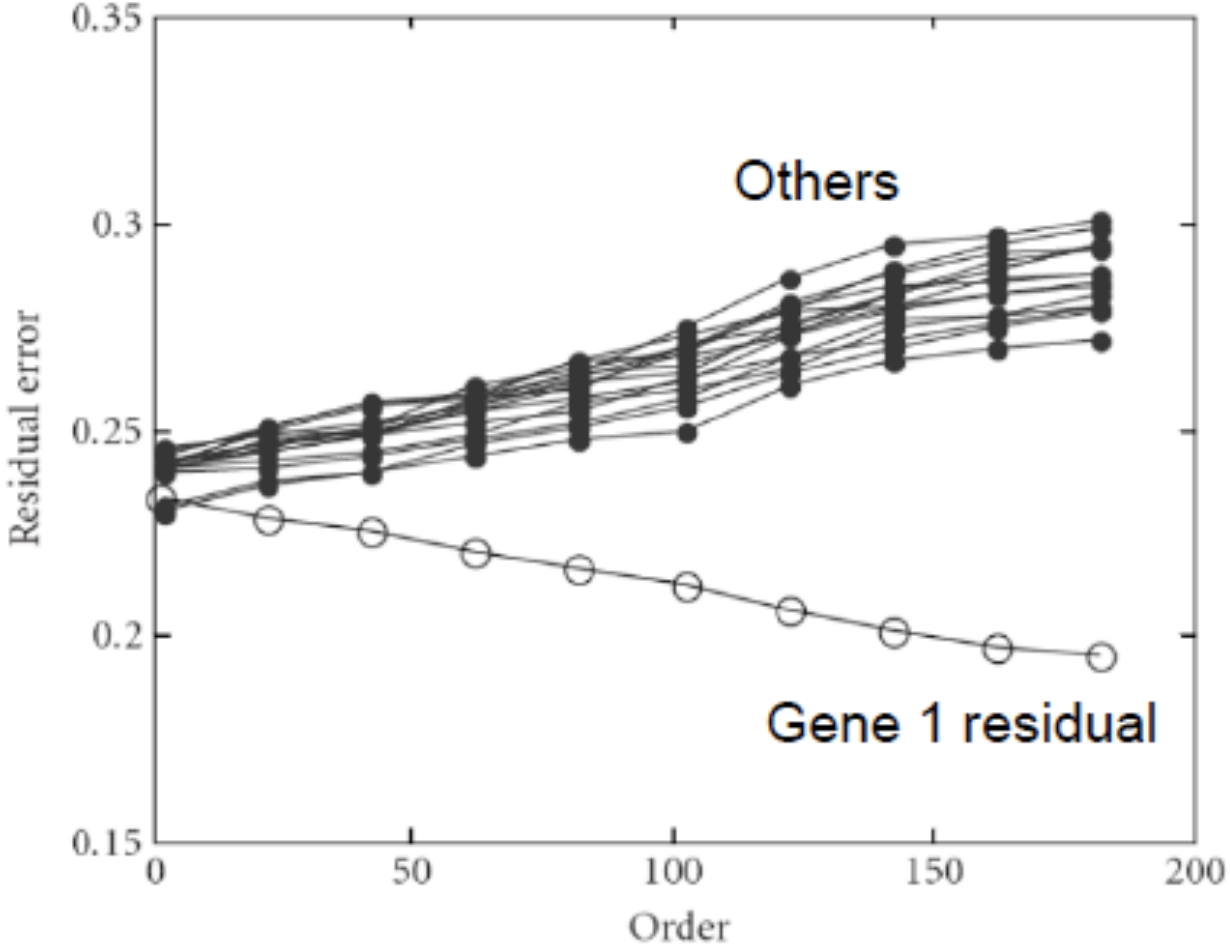
Advantage: Different Length DNA -- get comparable parameters(distance and correlations)

Disadvantage: Need high-order models? (Speech ~ order of 8 to 10 coeffs)

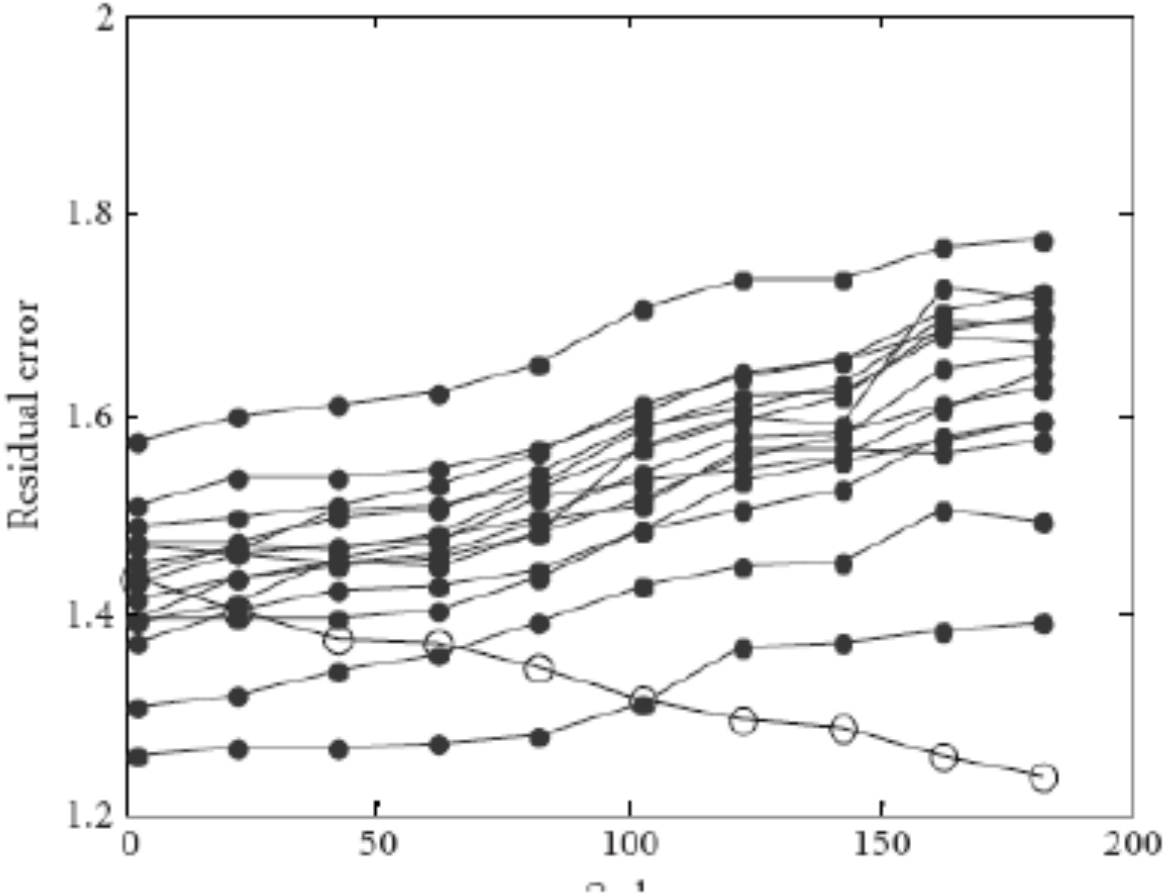
Analysis 3

- Says that for comparing spectra, need high order models

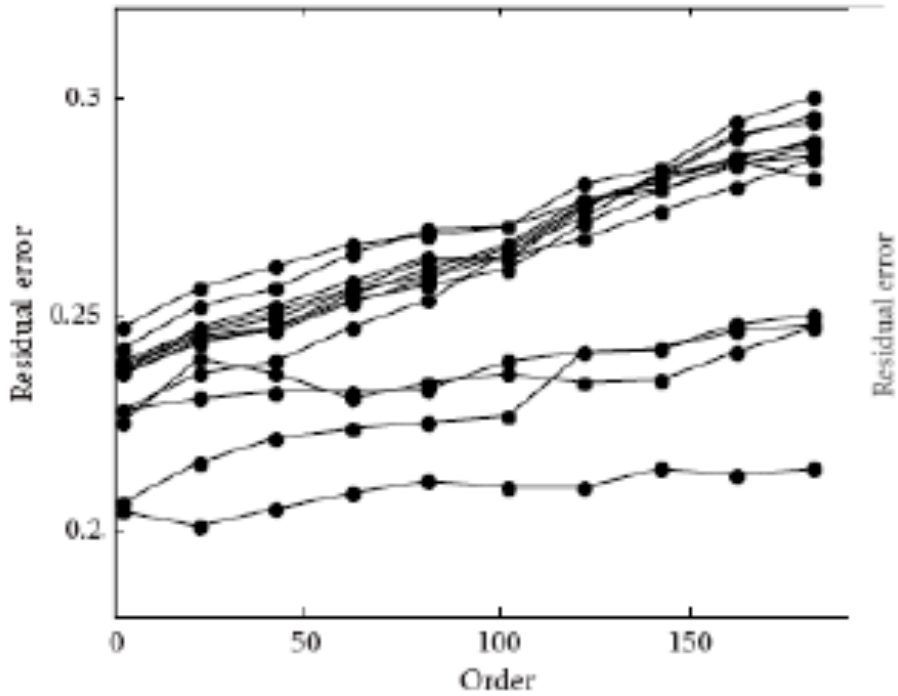
Residual from Gene1 AR model (binary indicator)



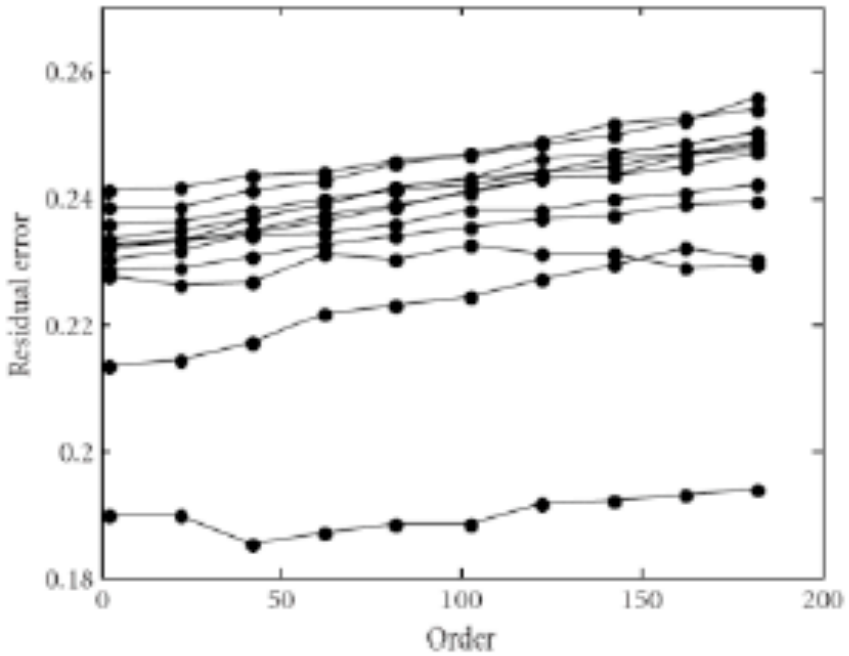
Residual from Gene1 AR model (Real-number)



AR Gene models with noncoding



Gene 1 with some noncoding seqs



Gene 17 with 36-50 noncoding

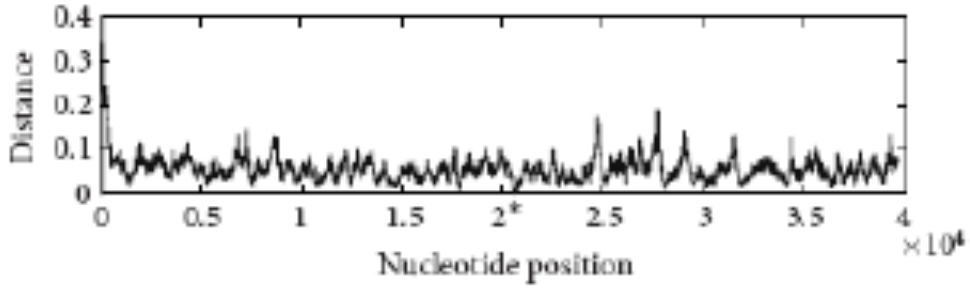
Models a noncoding one better than itself

Models another better than itself

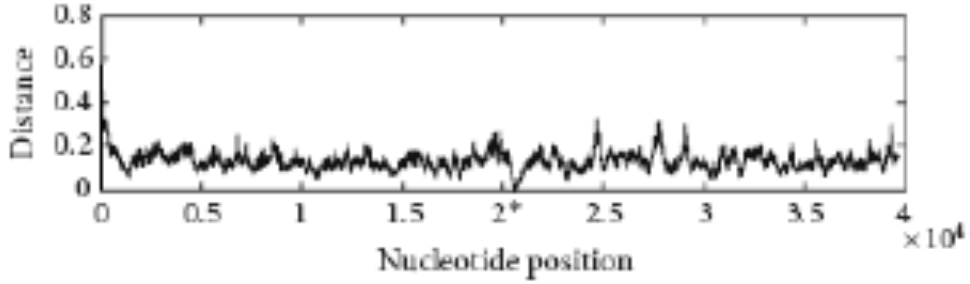
Moving algorithm

1. Calculate AR parameters for a template
2. Calculate AR parameters for a window length, L, of nucleotides
3. Calculate Euclidean distance between feature vectors
4. Increment by a small bit (overlapping windows)
5. Repeat 2 through 5

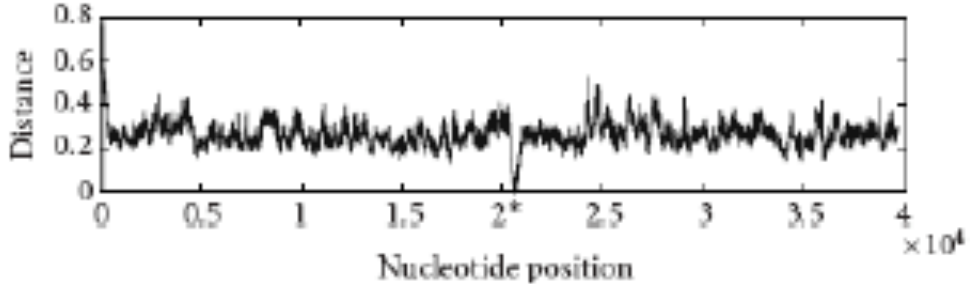
Distance between feature vectors



(a)



(b)



Itakura Distance

- How much better is **a** in predicting Frame 1 than **b?**
- $d(\mathbf{a},\mathbf{b}) = \log(E_b / E_a)$
- How much better is **a** in predicting Frame 1 than **b?**
- Not symmetrical so use:

$d_{avg}(\mathbf{a},\mathbf{b}) = 1/2[d(\mathbf{a},\mathbf{b})+d(\mathbf{b},\mathbf{a})]$

Homework

- Major differences in nucleotide biases:

Dictyostelium firmibasis plasmid Dfp1, NC_001923

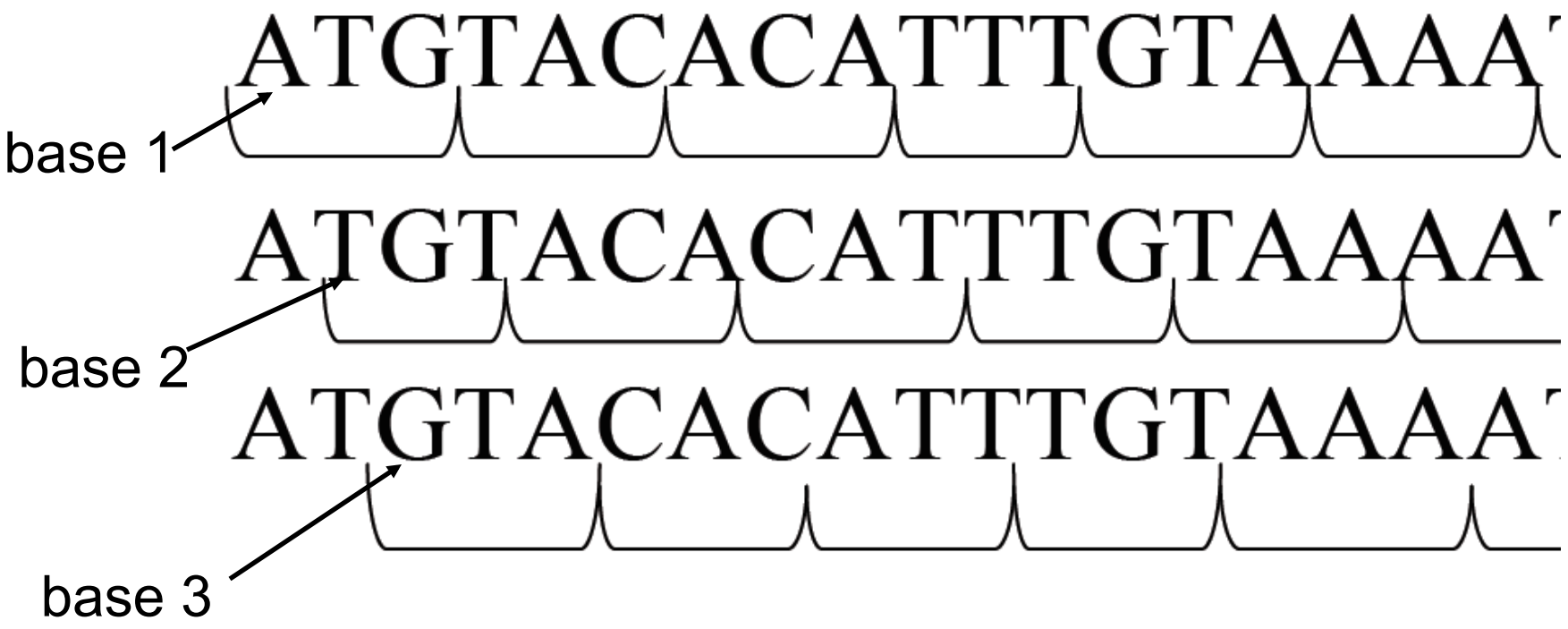
A: 2053
C: 567
G: 634
T: 1761

76% CG Content

```
>> codoncount(Dfp1)
AAA - 152    AAC -  31    AAG -  27    AAT -  74
ACA -  57    ACC -  15    ACG -   2    ACT -  32
AGA -  41    AGC -   8    AGG -   5    AGT -  36
ATA -  84    ATC -   8    ATG -  34    ATT -  77
CAA -  25    CAC -   5    CAG -   4    CAT -  23
CCA -  25    CCC -   2    CCG -   4    CCT -   9
CGA -   9    CGC -   0    CGG -   0    CGT -   8
CTA -  20    CTC -   2    CTG -   5    CTT -  26
GAA -  62    GAC -  13    GAG -  14    GAT -  68
GCA -  22    GCC -  10    GCG -   0    GCT -   2
GGA -   9    GGC -   4    GGG -   2    GGT -  17
GTA -  31    GTC -   4    GTG -   5    GTT -  38
TAA -  40    TAC -  18    TAG -  13    TAT -  83
TCA -  48    TCC -   6    TCG -   3    TCT -  16
TGA -  13    TGC -   1    TGG -   8    TGT -  25
TTA -  79    TTC -  20    TTG -  21    TTT - 126
```

Open Reading Frame Review

Any given nucleotide sequence (single DNA strand or mRNA) can be interpreted in three possible ways, depending on where the coding starts.



Base count for each base position

- Elegant Code
  - x1=x(1:3:end);
    basecount(x1);
  - x2=x(2:3:end);
    basecount(x2);

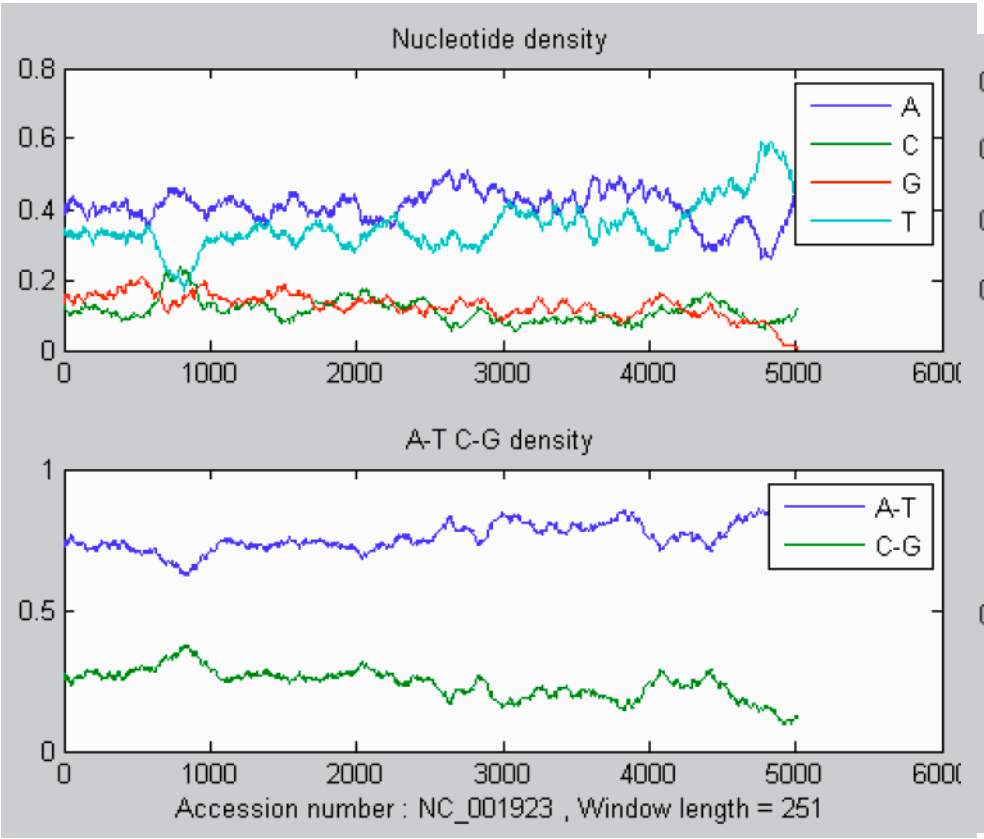
Human Enterovirus C

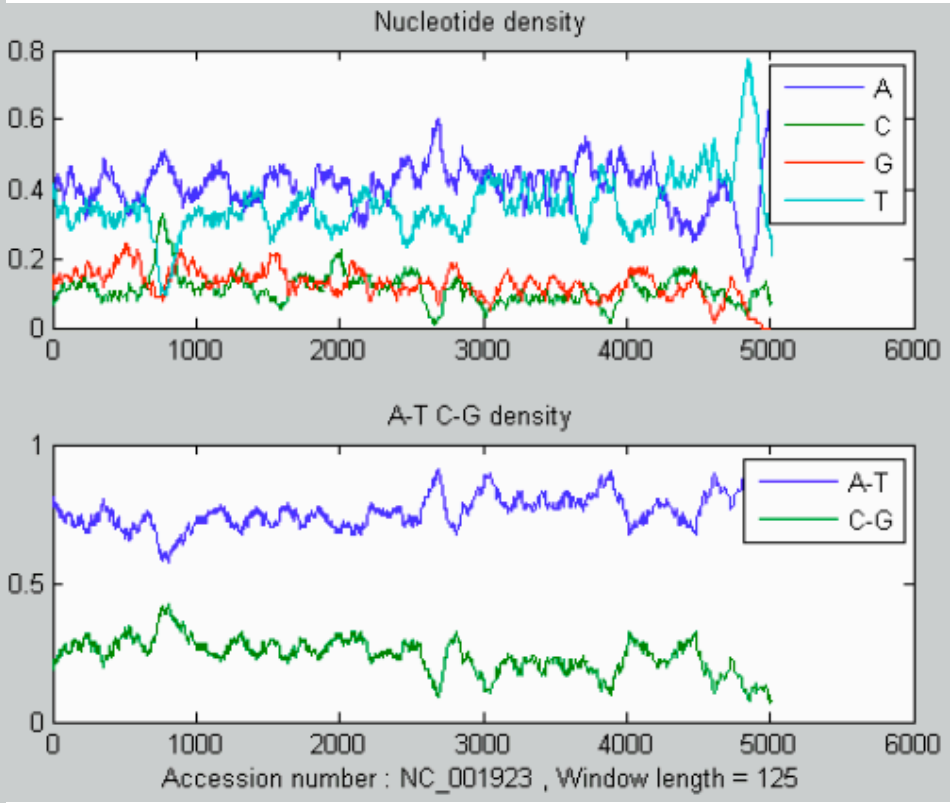
A	C	G	T
781	437	738	511
731	614	443	679
682	596	513	676

Dfp1

A	C	G	T
683	167	301	521
653	253	186	580
717	147	147	660

Window Differences





GC-rich / GC-poor

- http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=152811 (Substitution Pressure is AT-biased)
- http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1463024 (GC Rich gene produces 10x as much protein as poor one)