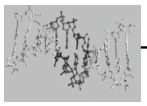
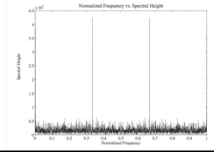
# Frequency Analysis of the Genome

Professor Gail L. Rosen



- ✓G: 1 0 0 0 0 0 0 1 0 0 0 1 0 0
- ✓C: 0 1 0 0 0 0 0 0 1 0 0 0 0 0
- ✓A: 0 0 1 0 0 1 0 0 0 1 1 0 0 0
- ✓T: 0 0 0 1 1 0 1 0 0 0 0 0 1 1




## Nucleotides are parts/code

- ✓How would you start to analyze it?
- ✓How would you represent each nucleotide in mathematical notation?

## Simple Representation

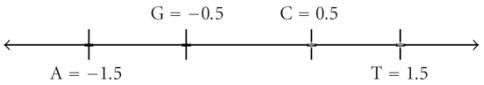
- ✓Arbitrary nucleotide assignment to integers
  - ✓ A<->0
  - ✓ C<->1
  - ✓ G<->2
  - ✓ T<->3
- ✓Does this make sense?

## Binary Indicator Sequence

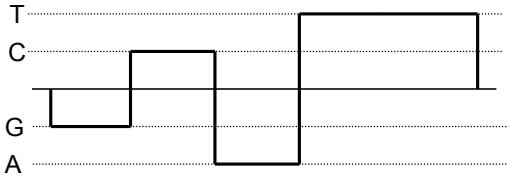
- ✓G C A T T A T G C A A G T T
  - ✓G: 1 0 0 0 0 0 0 1 0 0 0 1 0 0
  - ✓C: 0 1 0 0 0 0 0 0 1 0 0 0 0 0
  - ✓A: 0 0 1 0 0 1 0 0 0 1 1 0 0 0
  - ✓T: 0 0 0 1 1 0 1 0 0 0 0 0 1 1
- ✓What is sequence representation useful for?

## Amplitude (Pulse Amplitude Modulation in Telecom)

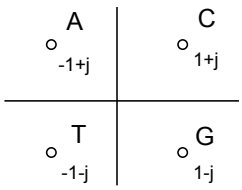
✓ Negation ~ Complement




GCATT




## Complex Representation (similar QPSK in Telecom)



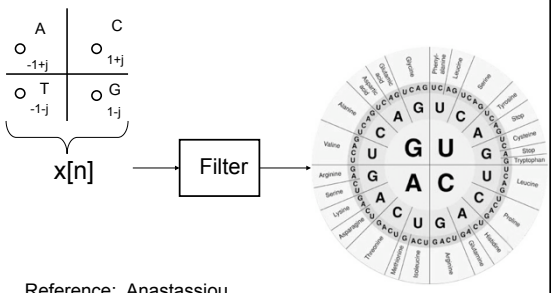

01 - C
00 - A
10 - T
11 - G

Rotation:

$a = -1,$
$c = -j,$
$g = j,$
$t = 1.$

✓ Conjugate ~ Complement

## Application: Nucleotide->Amino Acid "filtering"



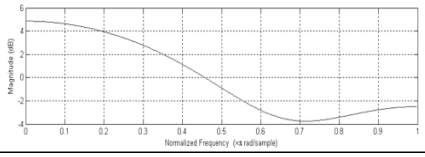

Reference: Anastassiou

## Filter coefficients in paper

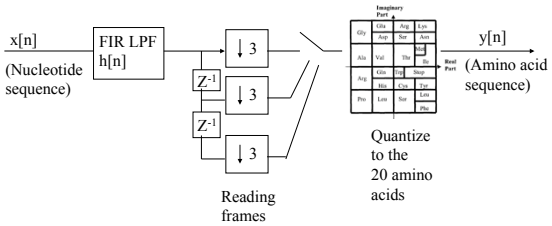
$$y[n] = h[0]\underbrace{x[n]}_{\text{base 1}} + h[1]\underbrace{x[n-1]}_{\text{base 2}} + h[2]\underbrace{x[n-2]}_{\text{base 3}}$$

Arbitrary filter coeffs

$h[0] = 1,\ h[1] = 1/2,\ h[2] = 1/4$




## For all Open Reading Frames

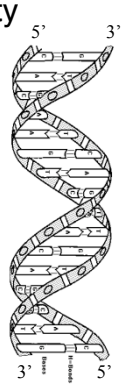



## Reverse complement property

- Ribosomes read from 5' to 3' end

Reverse complement of GCATT is AATGC

| 5' 3' |
|---|
| G - C |
| C - G |
| A - T |
| T - A |
| T - A |
| 3' 5' |




## Complex representation: reverse complement and conjugate symmetric

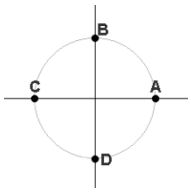
$$\tilde{x}[n] = x^*[-n+N-1], \quad n = 0,1,\ldots,N-1$$

Conjugate symmetric, x[n], has a real fourier transform, linear phase, etc.

Review symmetry properties of Fourier Transform -- Schafer p. 55

## Quadrature Phase Shift Keying

### Transmitting Sines and Cosines




| Phase Change | Example State Change | Dibit |
|---|---|---|
| 0° | A->A | 00 |
| 90° | A->B | 01 |
| 180° | B->D | 11 |
| 270° | D->C | 10 |

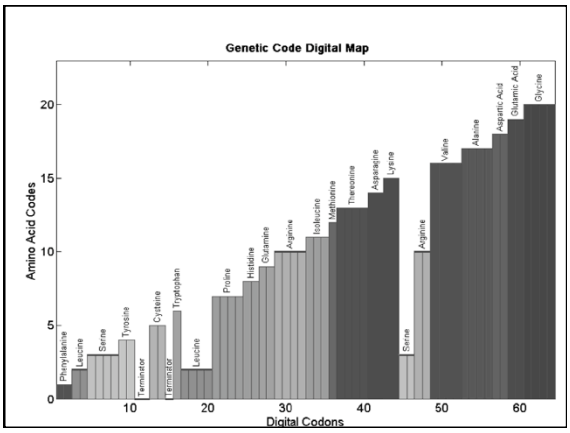
## Physical Ways of Representing

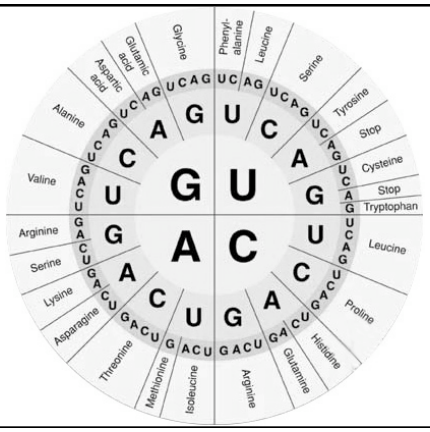
- ✓ Binary
  - ✓ Purine: A, G
  - ✓ Pyrimidine: T, C
- ✓ Electron-ion interaction potentials
- ✓ Bonding affinities

**Table 3. Mapping of Nucleotides to Digits in Base Four**

| | |
|---|---|
| ***Pyrimidines*** | |
| Thymine | = **T** = 0 |
| Cytosine | = **C** = 1 |
| ***Purines*** | |
| Adenine | = **A** = 2 |
| Guanine | = **G** = 3 |

| Nucleotide | EIIP |
|---|---|
| A | 0.1260 |
| G | 0.0806 |
| C | 0.1340 |
| T | 0.1335 |




## Human Coding Regions (Nucleotide ORF bias)

| nucleotide | codon position 1 | 2 | 3 |
|---|---|---|---|
| A | 0.27 | 0.31 | 0.18 |
| C | 0.24 | 0.24 | 0.31 |
| G | 0.32 | 0.20 | 0.29 |
| T | 0.17 | 0.26 | 0.22 |
| | C/G Pref | A/T Pref | C/GPref |

## Human Codon Usage

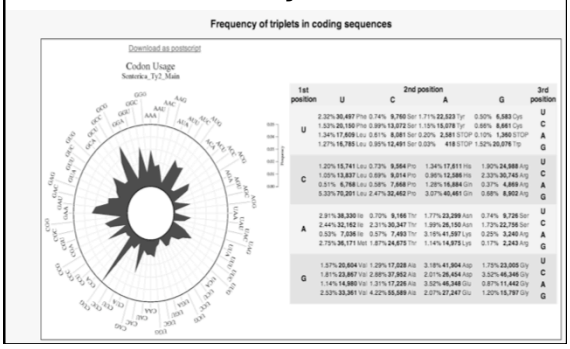
Usage of each codon per 1000 in coding regions

Percentage of codon composition among synonomous codons

The Human Codon Usage Table

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly | GGG | 17.08 | 0.23 | Arg | AGG | 12.09 | 0.22 | Trp | TGG | 14.74 | 1.00 | Arg | CGG | 10.40 | 0.19 |
| Gly | GGA | 19.31 | 0.26 | Arg | AGA | 11.73 | 0.21 | End | TGA | 2.64 | 0.61 | Arg | CGA | 5.63 | 0.10 |
| Gly | GGT | 13.66 | 0.18 | Ser | AGT | 10.18 | 0.14 | Cys | TGT | 9.99 | 0.42 | Arg | CGT | 5.16 | 0.09 |
| Gly | GGC | 24.94 | 0.33 | Ser | AGC | 18.54 | 0.25 | Cys | TGC | 13.86 | 0.58 | Arg | CGC | 10.82 | 0.19 |
| Glu | GAG | 38.82 | 0.59 | Lys | AAG | 33.79 | 0.60 | End | TAG | 0.73 | 0.17 | Gln | CAG | 32.95 | 0.73 |
| Glu | GAA | 27.51 | 0.41 | Lys | AAA | 22.32 | 0.40 | End | TAA | 0.95 | 0.22 | Gln | CAA | 11.94 | 0.27 |
| Asp | GAT | 21.45 | 0.44 | Asn | AAT | 16.43 | 0.44 | Tyr | TAT | 11.80 | 0.42 | His | CAT | 9.56 | 0.41 |
| Asp | GAC | 27.06 | 0.56 | Asn | AAC | 21.30 | 0.56 | Tyr | TAC | 16.48 | 0.58 | His | CAC | 14.00 | 0.59 |
| Val | GTG | 28.60 | 0.48 | Met | ATG | 21.86 | 1.00 | Leu | TTG | 11.43 | 0.12 | Leu | CTG | 39.93 | 0.43 |
| Val | GTA | 6.09 | 0.10 | Ile | ATA | 6.05 | 0.14 | Leu | TTA | 5.55 | 0.06 | Leu | CTA | 6.42 | 0.07 |
| Val | GTT | 10.30 | 0.17 | Ile | ATT | 15.03 | 0.35 | Phe | TTT | 15.36 | 0.43 | Leu | CTT | 11.24 | 0.12 |
| Val | GTC | 15.01 | 0.25 | Ile | ATC | 22.47 | 0.52 | Phe | TTC | 20.72 | 0.57 | Leu | CTC | 19.14 | 0.20 |
| Ala | GCG | 7.27 | 0.10 | Thr | ACG | 6.80 | 0.12 | Ser | TCG | 4.38 | 0.06 | Pro | CCG | 7.02 | 0.11 |
| Ala | GCA | 15.50 | 0.22 | Thr | ACA | 15.04 | 0.27 | Ser | TCA | 10.96 | 0.15 | Pro | CCA | 17.11 | 0.27 |
| Ala | GCT | 20.23 | 0.28 | Thr | ACT | 13.24 | 0.23 | Ser | TCT | 13.51 | 0.18 | Pro | CCT | 18.03 | 0.29 |
| Ala | GCC | 28.43 | 0.40 | Thr | ACC | 21.52 | 0.38 | Ser | TCC | 17.37 | 0.23 | Pro | CCC | 20.51 | 0.33 |

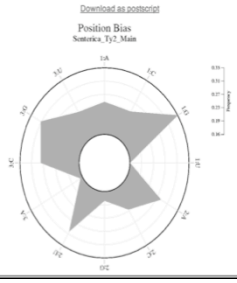
## Codon Usage in *Salmonella enterica* strain Ty2



## SE2 Ty2

**Nucleotide bias relative to position**

Download as postscript



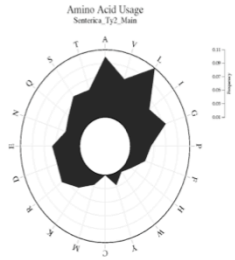

**Sequence Logo of Shannon information**

| Base | 1st position | | 2nd position | | 3rd position | |
|---|---|---|---|---|---|---|
| U | 15.63% | **205,725** | 30.24% | **398,099** | 24.33% | **320,329** |
| C | 24.48% | **322,301** | 22.91% | **301,588** | 28.04% | **369,177** |
| A | 25.02% | **329,366** | 28.57% | **376,116** | 16.34% | **215,182** |
| G | 34.88% | **459,150** | 18.29% | **240,739** | 31.28% | **411,854** |

## SE2 Ty2

**Amino Acid usage**

Download as postscript




| Letter | Amino Acid | Frequency | Count | Group |
|---|---|---|---|---|
| A | A | 9.74% | **127,844** | Aliphatic |
| V | V | 7.04% | **92,455** | Aliphatic |
| L | L | 10.73% | **140,862** | Aliphatic |
| I | I | 5.91% | **77,559** | Aliphatic |
| G | G | 7.36% | **96,580** | Structural |
| P | P | 4.47% | **58,710** | Structural |
| F | F | 3.86% | **50,653** | Aromatic |
| H | H | 2.30% | **30,199** | Aromatic |
| W | W | 1.53% | **20,081** | Aromatic |
| Y | Y | 2.86% | **37,598** | Aromatic |
| C | C | 1.16% | **15,237** | Sulfur |
| M | M | 2.80% | **36,718** | Sulfur |
| K | K | 4.31% | **56,591** | + |
| R | R | 5.71% | **74,947** | + |
| D | D | 5.21% | **68,392** | - |
| E | E | 5.61% | **73,632** | - |
| N | N | 3.77% | **49,470** | Polar |
| Q | Q | 4.37% | **57,357** | Polar |
| S | S | 5.78% | **75,891** | Polar |
| T | T | 5.46% | **71,707** | Polar |

## Fourier Transform

Another: Define $u_A[n]$, $u_T[n]$, $u_C[n]$, $u_G[n]$ as binary indicator sequences

A A G T G C $\Longleftrightarrow$

| | | | | | | |
|---|---|---|---|---|---|---|
| $u_A[n]$ | 1 | 1 | 0 | 0 | 0 | 0 |
| $u_G[n]$ | 0 | 0 | 1 | 0 | 1 | 0 |
| $u_T[n]$ | 0 | 0 | 0 | 0 | 0 | 1 |
| $u_C[n]$ | 0 | 0 | 0 | 1 | 0 | 0 |

frequency sequence →

$$U_A[k] = \sum_{n=1}^{N} u_A[n] * e^{-jkn}$$

Index sequence

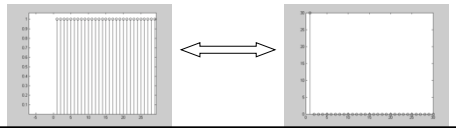
## Transform of the whole sequence
(modifications for binary indicator rep)

$$X[k] = aU_A[k] + tU_T[k] + cU_C[k] + gU_G[k]$$
$$k = 0,1,\ldots,N-1$$

Why not just take *a=t=c=g*?

- $x[n] = u_A[n] + u_T[n] + u_C[n] + u_G[n] = [1\ 1\ 1\ 1\ \ldots.\ 1]$ (no information)

x[n] X[k]

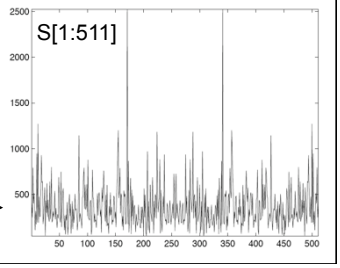


## Spectral Content

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2$$

Example mitochondrial sequence (from seqstatsdemo Took N=512 point DFT). Spikes at k=N/3 and 2N/3 are the same.

S[0], the dc term, is subtracted out →




## Recap

### Periodicity in DNA Structure

- Codons that code for specific amino acids are 3 bases in length.
  - Open Reading Frame (ORF)
- DNA coding sequences exhibit 3-base periodicity
- DNA non coding sequence exhibit no periodicity

## Review: Open Reading Frame

Frame Offset

0 ATGTACACATTTGTAAAATGA

1 ATGTACACATTTGTAAAATGA

2 ATGTACACATTTGTAAAATGA

✓ Periodicities occur in Codon Position

## Reason for Periodicity in DNA

- Imbalance in distribution of nucleotides in each ORF position
  - Caused by protein preference towards certain amino acid combinations
  - Bias in coding region that does not exist in non-coding regions.

## Processing a DNA Sequence

1. Acquire DNA Sequence
2. Transform the character string into a numeric representation
3. Transform numeric string into the Frequency Domain
4. Check for a peak at frequency, f=1/3

IMPORTANT HINT: Remove the DC Component when plotting

## Numeric Representation: BIS

- Binary Indicator Sequences (BIS)
- Parse through the DNA sequence and mark all locations with a specific base as a 1, all others as a 0

## Discrete Fourier Transform (DFT)

- Way to transform data into the frequency (aka Fourier) domain
- Requires input function that is:
  - Discrete
  - Finite Duration

## Calculate the DFT

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, ..., N-1$$

- e is The exponential function
- i is the imaginary unit
- N is the length of the DFT
- In order for the DFT to have full resolution and not truncate data $N \geq M$ ,M=length of the original sequence

## Apply DFT to BIS

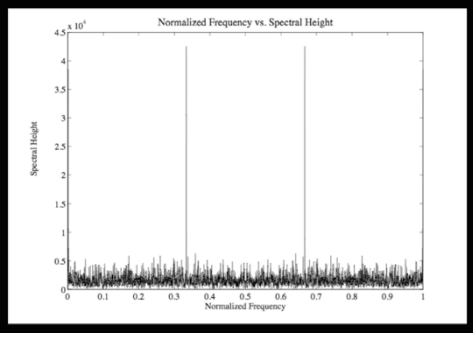
- Take the DFT of each BIS

$$U_A = \sum_{n=0}^{N-1} u_A(n) e^{\frac{-2\pi i}{N} kn}$$

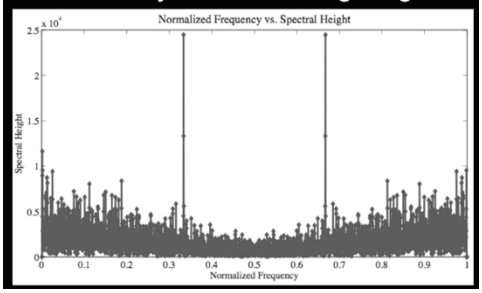
- To plot spectrum of DNA sequence sum the squares of the DFTs of all BIS

$$DNASpec = |U_A|^2 + |U_C|^2 + |U_G|^2 + |U_T|^2$$

## DNA Sequence in the Fourier Domain: Coding Region of E. Coli

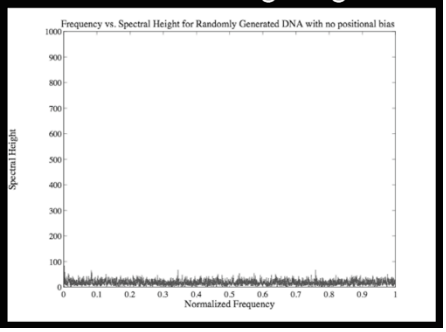



## DNA Sequence in the Fourier Domain: Synthetic Coding Region

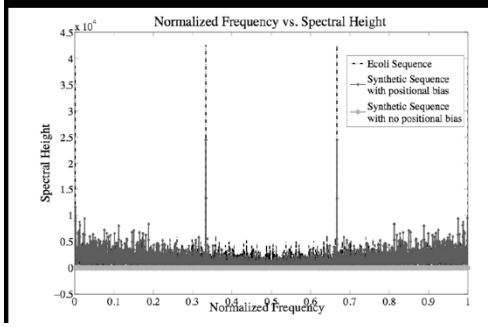


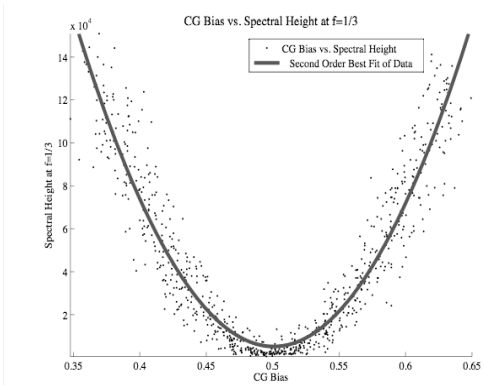

Only one base biased

## DNA Sequence in the Fourier Domain: Non Coding Region

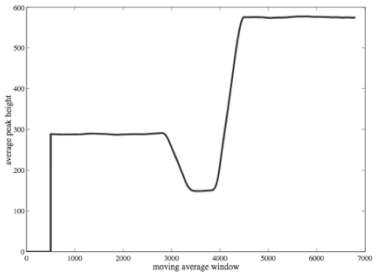



## DNA Sequence in the Fourier Domain

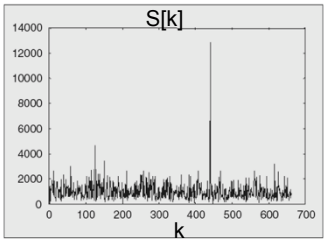


## Can Use Height to Detect Different Coding Regions




## Anastassiou's example




▲ 9. Plot of the spectrum of a coding DNA region, demonstrating peak at frequency $k = N/3$.

- Baker's Yeast -- sequence and DFT length of 1320
- Usually frequency axis is 0 -> $\pi$.
- frequency ~ 1/ wavelength
- k ~ 1/period-length
- k= N / (periodicity-length)

## Yin/Yau - Background Noise

✓ Noise[k]= S[k]/Seq_length
(Average Power over every frequency)

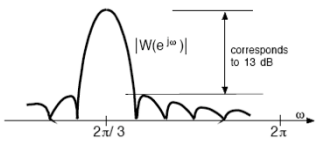
## Why is there a period of 3?

✓ If each base equiprobable, no period
✓ CG, codon bias
✓ Abundance of G in position 1
✓ Tiwari et al. "synthesized" genes backwards and found period-3
✓ Tiwari et al. found that some genes in S. Cerevisiae do not have period-3

## Effects of using a "sliding" DFT window

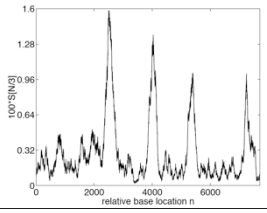
Fine tune frequency

$$w(n) = \begin{cases} e^{j\omega_0 n} & 0 \le n \le N-1 \\ 0 & \text{otherwise.} \end{cases}$$




## Gene prediction using DFT sliding window

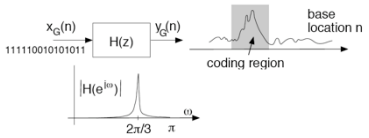
- Plot S[N/3] as a function of a moving window
  - What is the window length?
  - What is the overlap of the windows?




F56F11.4 in the C-elegans chromosome III

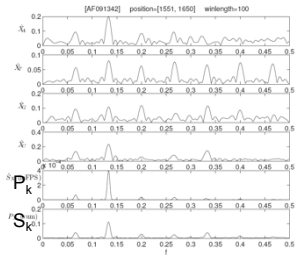
## "Improved filtering" for gene prediction

- ✓ If get peak at N/3, coding region
- ✓ Vaidyanathan and Yoon
  - ✓ Anti-Notch Filtering




## Fourier Product Spectrum, P[k]

- ✓ Multiply spectrums together
- ✓ $P[k] = |U_A[k]| * |U_C[k]| * |U_G[k]| * |U_T[k]|$
- ✓ Amplifies peaks




## Issues with the Spectral methods

- ✓ Can we exploit the spectrum to also signify structural attributes of the sequence?
- ✓ Why just the magnitude? Is there no phase information to exploit? Assume that a lot of information from coding to non-coding (frameshifts).

## Coding Bias Measure from Spectrums (Yin/Yau 2005)

Occurence of each nucleotide in each ORF position for nucleotide x:

$$F_{x1} \quad F_{x2} \quad F_{x3}$$

The spectral peak height to these occurences

$$PS(N/3) = \sum_{x=A,T,C,G} [F_{x1}^2 + F_{x2}^2 + F_{x3}^2 - (F_{x1} * F_{x2} + F_{x1} * F_{x3} + F_{x2} * F_{x3})].$$

## $X_A[N/3]$ ~ Pr(A in ORF1 ∪ A in ORF2 ∪ A in ORF3)

- ✓ Measure of how frequent A is every 3 nucleotides

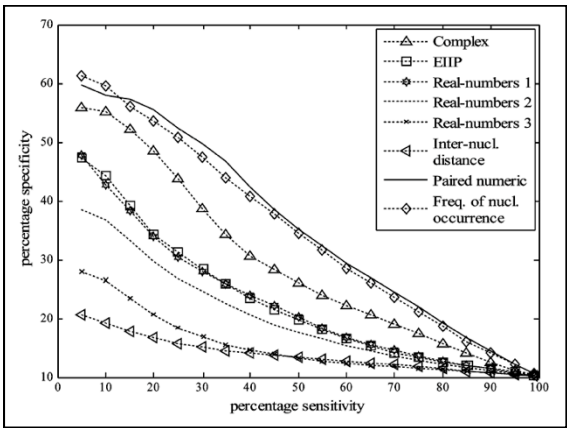
Frame Offset

| | |
|---|---|
| 0 | ATGTACACATTTGTAAAATGA |
| 1 | ATGTACACATTTGTAAAATGA |
| 2 | ATGTACACATTTGTAAAATGA |

ORF 1 Sequence: ATATGAT $F_{A1}$=3/7
ORF 2 Sequence: TACTTAG $F_{A2}$=2/7
ORF 3 Sequence: GCATAAA $F_{A3}$=4/7

## Mahmood and Epps: Numeric Representation can affect DFT

- Complex
- EIIP (electron-ion interaction potential)
- Real Numbers
  - T=0; C=1; A=2;G=3
  - A=0; G=1; C=2; T=3
  - A=1.5; T=-1.5, C=0.5,G=0.5 (Amplitude Modulation)
- Internucleotide Difference (replaces each DNA nucleotide with an integer representing the distance between the current nucleotide and the next similar nucleotide.)
- Paired Numeric (A-T: 1, C-G:0)
- Frequency of Nucleotide Occurrence




| Period-3 Detection Method | Data-driven (Y/N) | Burset/Guigo1996 | | | | | HMR195 | | | | | GENSCAN test set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Area under ROC curve | % impr. Over SC | % of exonic nucleotides detected at false positive | | | Area under ROC curve | % impr. Over SC | % of exonic nucleotides detected at false positive | | | Area under ROC curve | % impr. Over SC | % of exonic nucleotides detected at false positive | | |
| | | | | 10% | 20% | 30% | | | 10% | 20% | 30% | | | 10% | 20% | 30% |
| SC | N | 0.7634 | — | 42.9 | 59.8 | 70.5 | 0.8008 | — | 49.1 | 65.0 | 75.0 | 0.7778 | — | 46.7 | 61.6 | 71.0 |
| SR | Y | — | — | — | — | — | — | — | — | — | — | 0.7800 | 0.29 | 48.6 | 62.9 | 72.4 |
| PWSR | Y | — | — | — | — | — | — | — | — | — | — | 0.8123 | 4.44 | 53.8 | 68.7 | 77.3 |
| PSC | N | 0.7702 | 0.88 | 46.2 | 61.0 | 70.7 | 0.8061 | 0.66 | 52.0 | 66.9 | 75.8 | 0.8114 | 4.32 | 52.3 | 68.3 | 77.6 |
| ACF | N | 0.5795 | −24.09 | 15.5 | 29.4 | 41.4 | 0.6340 | −20.83 | 20.3 | 35.1 | 47.9 | 0.6218 | −20.06 | 18.4 | 33.3 | 47.1 |
| AMDF | N | **0.8065** | **5.64** | **52.2** | **67.2** | **76.4** | **0.8323** | **3.93** | 55.1 | 70.5 | **79.7** | 0.8338 | 7.20 | 56.2 | 72.9 | 81.7 |
| TDP | N | 0.7876 | 3.17 | 50.5 | 63.8 | 72.6 | 0.8258 | 3.12 | **57.5** | **70.6** | 78.0 | 0.8335 | 7.16 | 59.9 | 72.5 | 79.6 |
| AR | N | 0.6632 | −13.13 | 29.0 | 43.5 | 54.4 | 0.7128 | −11 | 34.1 | 50.0 | 61.3 | 0.7057 | −9.27 | 35.9 | 51.7 | 63.2 |
| AN filter | N | 0.6735 | −11.78 | 31.5 | 45.6 | 56.0 | 0.7118 | −11.12 | 37.1 | 51.2 | 61.3 | 0.6875 | −11.60 | 32.5 | 47.6 | 58.1 |
| SVD | N | 0.7729 | 1.24 | 48.0 | 61.8 | 70.6 | 0.8152 | 1.79 | 54.8 | 68.2 | 76.6 | 0.8252 | 6.09 | 58.7 | 70.9 | 78.5 |
| TFH | Y | — | — | — | — | — | — | — | — | — | — | 0.8448 | 8.62 | 59.5 | 74.9 | 81.6 |
| TFH (5D) | Y | — | — | — | — | — | — | — | — | — | — | 0.8542 | 9.82 | 62.3 | 76.2 | 83.7 |
| AR–TFH (17D) | Y | — | — | — | — | — | — | — | — | — | — | **0.8739** | **12.36** | **66.3** | **80.5** | **87.0** |

$$\mathrm{AMDF}[k] = \frac{1}{N}\sum_{n=1}^{N} |x[n] - x[n-k]|$$

✓AR-TFH: AR parameters + Time-Frequency parameters (magnitude +phase)

TABLE I
SUMMARY OF DATASETS

| Dataset | Organisms | # gene sequence | # bp | # exon | Coding density (%) |
|---|---|---|---|---|---|
| Burset/Guigo1996 [10] | Vertebrate | 570 | 2,892,149 | 2649 | 15.37 |
| HMR195 [11] | Mammalian | 195 | 1,383,720 | 948 | 14 |
| GENSCAN Learning Set [45] | Human | 380 | 2,581,000 | 1492 | 16 |
| GENSCAN Test Set [45] | Human | 65 | 591,886 | 381 | 10.2 |