Signal Processing Problems in Genomics

P. P. Vaidyanathan
California Institute of Technology
Pasadena, CA
Why is genomics interesting for the signal processing person?

Because there are sequences there!

OK, what sort of sequences?

1. Sequences from an alphabet of size four:
   ...ATTCGAAGATTTTCAACCGGGAAAAA...
   DNA

2. Sequences from an alphabet of size twenty:
   AACWYDEFGHIKLMNPQRSTVAPPQR
   Protein

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Size-4 alphabet:
A, C, T, G: bases (also called or nucleotides)

DNA sequences (genomes) are made of these.

Genes are parts of DNA, and are 4-letter sequences.

Adenine  Thymine  Cytosine  Guanine
or Uracil (in RNA)

DNA: deoxyribonucleic acid
RNA: ribonucleic acid

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
DNA molecule in the living cell (usually in nucleus)

Complementary Strands in the Double Helix

A  T
C  G

Great place to get started, and a great reference

Alberts, et. al., Essential Cell Biology, Garland publishing, Inc., 1998

Alberts, Bray, Johnson, Lewis, Raff, Roberts, and Walter
A good introductory article (signal processing aspects)
Size-20 alphabet:

ACDEFGHIKLMNPQRSTVWY: **amino acids**

*(B,J,O,U,X,Z missing)*

**Proteins are sequences made of these letters.**

20-letter proteins and 4-letter DNA are common to all life

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
The twenty natural amino acids
(B,J,O,U,X,Z missing)

11 essential amino acids.

Animals cannot make the eleven indicated amino acids.
They need to eat them;

Milk provides all of these.
Grains and beans together provide all of these.

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Protein Example

Fibroblast growth factor proteins

Basic bovine
PALPEDGGSGAFPPGHFKDPKRLYCKNGGF
FLRIHPDGRVDGVREKSDPHIKLQLQAEER
GVVSIKGVCANYLAMKEDGRLLLASKCVTD
ECFFFERLESNNYNTYRSRKYSSWYVALKR
TGQYKLPKTPGQPQKAIIFLPLMSAKS

Acidic bovine
FNPLGLNYKKPKLLYCSNGGYFLRILPDGT
VDGTKDRSDQHIQLQLCAESIGEYVYIKSTE
TGQFLAMDTDGLLYGTSQTPNEECLFLERLE
ENHYNTYISKKHAEKHWFVGLKKNGRSKLG
PRTHFGQKAILFLPLPVSSD

Will return to these and talk about their Fourier transforms

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Outline

• Molecular biology background

• Computational gene-finding

• Spectral analysis (Fourier, wavelet, correlations)

• Hidden Markov Models and sequence analysis

• New world of non-coding genes

• References

Will try to cover the cream of it.

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Bacterial DNA: few \textbf{million} bases;  Human DNA: three \textbf{billion} bases

\textbf{If we write the bases as letter-sized objects:}
- Bacterial DNA takes up the space of about 50 average novels.
- Human DNA takes about 2000 novels.

\textbf{Actual physical size:}
- human DNA in any cell stretches out to \textbf{2 yards}.
- DNA in all 5 trillion cells in humans:

\begin{center}
\textit{Covers it 50 times over}
\end{center}

\textit{P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver}
What do genes do?

Top strand, let’s say

Protein 1

Protein 2

Protein 3

DNA sequence

1

2

3

Intergenic spaces; contain A,T,C,G too!

Lots of protein in the cell, inside and outside nucleus

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
All cells in a given organism have the same DNA; same set of genes. But **different genes are expressed** (i.e., functional) in different cells. That’s why **brain cells** look different from **blood cells**, and so forth.

When a **gene** is **expressed**, it gives instructions to the cell to make a particular **protein**.

*Each gene makes a different protein.*
Example of a Protein: Hemoglobin (oxy, human)

http://www.biochem.szote.u-szeged.hu/astrojan/protein2.htm

Sequence of amino acids. Folds into beautiful 3D shapes. Necessary for function.
Example of a protein (an enzyme)

http://www.biochem.szote.u-szeged.hu/astrojan/protein2.htm
some other molecule, e.g., ligand

protein molecule

Fits like a puzzle piece. That’s how beautifully enzymes work!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Generation of a protein from a gene

A,C,U,G sequence

Gene copied into mRNA (transcription)

reduced mRNA (introns removed by splicing)

Converted to protein by tRNA and ribosome

(translation from 4-language to 20-language)

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
In this process the ribosome works with a molecule called tRNA which transfers groups of 3 bases (codons) in the mRNA into amino acids that make up the protein. The protein folds beautifully into its 3D structure which depends only on the amino acid sequence (and pH of medium). Now it is ready to function.
Central dogma of molecular biology (Crick)

DNA $\xrightarrow{\text{transcript}}$ mRNA $\xrightarrow{\text{translate}}$ protein

Pioneers: Beadle and Tatum, Bread mold experiment (1942)

In recent years the central dogma has been challenged!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Role of codons

Gene from DNA scanned from 5’ to 3’ end:
5’ ATGGAAGTGGCAATGATCCTGAATTATAACGTACTAG 3’
The gene is interpreted in groups of three bases called **codons**.

<table>
<thead>
<tr>
<th>5’ end</th>
<th>3’ end</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGGAAGTGGCAATGATCCTGAATTATAACGTACTAG</td>
<td>gene</td>
</tr>
</tbody>
</table>

**ATG**: start codon; also codon for M (met); plays two roles

**TAA, TAG, TGA**: stop codons (*do not code for amino acids*).

*Typically genes are long (1000s of bases); proteins have 100s to 1000s of amin acids*

_P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver_
### The genetic code

<table>
<thead>
<tr>
<th>codon</th>
<th>amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA:</td>
<td>K (Lys)</td>
</tr>
<tr>
<td>AAG:</td>
<td>K (Lys)</td>
</tr>
<tr>
<td>AAT:</td>
<td>N (Asn)</td>
</tr>
<tr>
<td>AAC:</td>
<td>N (Asn)</td>
</tr>
<tr>
<td>AGA:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>AGG:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>AGT:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>AGC:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>ATA:</td>
<td>I (Ile)</td>
</tr>
<tr>
<td>ATG:</td>
<td>M (Met)</td>
</tr>
<tr>
<td>ATT:</td>
<td>I (Ile)</td>
</tr>
<tr>
<td>ATC:</td>
<td>I (Ile)</td>
</tr>
<tr>
<td>ACA:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACG:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACT:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACC:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>GAA:</td>
<td>E (Glu)</td>
</tr>
<tr>
<td>GAG:</td>
<td>E (Glu)</td>
</tr>
<tr>
<td>GAT:</td>
<td>D (Asp)</td>
</tr>
<tr>
<td>GAC:</td>
<td>D (Asp)</td>
</tr>
<tr>
<td>GGA:</td>
<td>G (Gly)</td>
</tr>
<tr>
<td>GGG:</td>
<td>G (Gly)</td>
</tr>
<tr>
<td>GGT:</td>
<td>G (Gly)</td>
</tr>
<tr>
<td>GCC:</td>
<td>G (Gly)</td>
</tr>
<tr>
<td>GTA:</td>
<td>V (Val)</td>
</tr>
<tr>
<td>GTG:</td>
<td>V (Val)</td>
</tr>
<tr>
<td>GTT:</td>
<td>V (Val)</td>
</tr>
<tr>
<td>GTC:</td>
<td>V (Val)</td>
</tr>
<tr>
<td>GCA:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCG:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCT:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCC:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>TAA:</td>
<td>STOP</td>
</tr>
<tr>
<td>TAG:</td>
<td>STOP</td>
</tr>
<tr>
<td>TAT:</td>
<td>Y (Tyr)</td>
</tr>
<tr>
<td>TAC:</td>
<td>Y (Tyr)</td>
</tr>
<tr>
<td>TGA:</td>
<td>STOP</td>
</tr>
<tr>
<td>TGG:</td>
<td>W (Trp)</td>
</tr>
<tr>
<td>TGT:</td>
<td>C (Cys)</td>
</tr>
<tr>
<td>TGC:</td>
<td>C (Cys)</td>
</tr>
<tr>
<td>TTA:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>TTG:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>TTT:</td>
<td>F (Phe)</td>
</tr>
<tr>
<td>TTC:</td>
<td>F (Phe)</td>
</tr>
<tr>
<td>TCA:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>TCG:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>TCT:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>TCC:</td>
<td>S (Ser)</td>
</tr>
<tr>
<td>CAA:</td>
<td>Q (Gln)</td>
</tr>
<tr>
<td>CAG:</td>
<td>Q (Gln)</td>
</tr>
<tr>
<td>CAT:</td>
<td>H (His)</td>
</tr>
<tr>
<td>CAC:</td>
<td>H (His)</td>
</tr>
<tr>
<td>CGA:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>CGG:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>CGT:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>CGC:</td>
<td>R (Arg)</td>
</tr>
<tr>
<td>CTA:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>CTG:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>CTT:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>CTC:</td>
<td>L (Leu)</td>
</tr>
<tr>
<td>ACA:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACG:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACT:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>ACC:</td>
<td>T (Thr)</td>
</tr>
<tr>
<td>GCA:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCG:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCT:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>GCC:</td>
<td>A (Ala)</td>
</tr>
<tr>
<td>CCA:</td>
<td>P (Pro)</td>
</tr>
<tr>
<td>CGC:</td>
<td>P (Pro)</td>
</tr>
<tr>
<td>CCT:</td>
<td>P (Pro)</td>
</tr>
<tr>
<td>CCC:</td>
<td>P (Pro)</td>
</tr>
</tbody>
</table>

The genetic code is common to ALL life!
Mutations in genes can cause disease

Gene HBB creates the protein beta globin in hemoglobin of red blood cells. This gene is 1600 bases long, and the spliced mRNA 626 bases long.

A single error in this sequence is responsible for sickle cell anemia.

DNA replicates itself when the cell divides.

AATATAGACCGACCCCTAAGTAAAAATAGACCTAGTAGA

1 error per billion bases.

\[ P_e = 10^{-9} \]

**Built-in proof reading system called mismatch-pair system**

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Parcel service, first class mail: 13 late deliveries out of 100 parcels
Airline luggage: 1 lost bag per 200
Professional typist: 1 mistake in 250 characters
Driving in the US: 1 death per 10,000 people per year
DNA replication: 1 error per billion bases copied
Speaker giving a talk: 1 error per slide
Beginning of the history of molecular biology:


http://www.pbs.org/wgbh/nova/photo51/before.html

End of this part
Outline

• Molecular biology background

• Computational gene-finding
  • Spectral analysis (Fourier, wavelet, correlations)
  • Hidden Markov Models and sequence analysis
  • New world of non-coding genes

• References

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
DNA AACTGGGCATCCCGGGAATAAGGTC

\[ x_A(n) = 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 \]

**Indicator sequence for base A**

Similarly define \( x_T(n), x_C(n), x_G(n) \)

\[ x_A(n) + x_T(n) + x_C(n) + x_G(n) = 1 \]

**Fourier transforms:**

\[ X_A(e^{j\theta}), X_T(e^{j\theta}), X_C(e^{j\theta}), X_G(e^{j\theta}) \]

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Fourier transforms:

\[ X_A(e^{j\omega}) \quad X_T(e^{j\omega}) \quad X_C(e^{j\omega}) \quad X_G(e^{j\omega}) \]

Define \( S(e^{j\omega}) \) to be the sum-of-magnitude squares.

In protein coding regions this exhibits a peak at \( 2\pi/3 \).

**Period-3 property.**

Even the plot of one base, e.g., \( X_G \) reveals this!

**Coding region of length \( N=1320 \) inside a genome of baker’s yeast (\( S. \) cerevisiae).**

Tiwari, et. al., CABIOS, 1997.
Period-3 property arises from the special bias built into the genetic code. Some bases dominate at certain positions, e.g., base G is dominant at positions 1 and 2.

<table>
<thead>
<tr>
<th>1</th>
<th>A</th>
<th>Ala</th>
<th>Alanine</th>
<th>GCA,GCC,GCG,GCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>C</td>
<td>Cys</td>
<td>Cysteine (has S)</td>
<td>TGC, TGT</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>Asp</td>
<td>Aspartic acid</td>
<td>GAC,GAT</td>
</tr>
<tr>
<td>4</td>
<td>E</td>
<td>Glu</td>
<td>Glutamic acid</td>
<td>GAA,GAG</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>Phe</td>
<td>Phenylalanine(^1)</td>
<td>TTC,TTT</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td>Gly</td>
<td>Glycine</td>
<td>GGA,GGC,GGG,GGT</td>
</tr>
<tr>
<td>7</td>
<td>H</td>
<td>His</td>
<td>Histidine(^2)</td>
<td>CAC,CAT</td>
</tr>
<tr>
<td>8</td>
<td>I</td>
<td>Ile</td>
<td>Isoleucine(^3)</td>
<td>ATA,ATC,ATT</td>
</tr>
<tr>
<td>9</td>
<td>K</td>
<td>Lys</td>
<td>Lysine(^4)</td>
<td>AAA,AAG</td>
</tr>
<tr>
<td>10</td>
<td>L</td>
<td>Leu</td>
<td>Leucine(^5)</td>
<td>TTA,TTG,CTA,CTC,CTG,CTT</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>Met</td>
<td>Methionine(^6) (has S)</td>
<td>ATG</td>
</tr>
<tr>
<td>12</td>
<td>N</td>
<td>Asn</td>
<td>Asparagine</td>
<td>AAC,AAT</td>
</tr>
<tr>
<td>13</td>
<td>P</td>
<td>Pro</td>
<td>Proline</td>
<td>CCA, CCC, CCG,CCT</td>
</tr>
<tr>
<td>14</td>
<td>Q</td>
<td>Gln</td>
<td>Glutamine</td>
<td>CAA,CAG</td>
</tr>
<tr>
<td>15</td>
<td>R</td>
<td>Arg</td>
<td>Arginine(^7)</td>
<td>AGA,AGG,CGA,CGC,CGG,CGT</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
<td>Ser</td>
<td>Serine</td>
<td>AGC,AGT,TCA,TCC,TCG,TCT</td>
</tr>
<tr>
<td>17</td>
<td>T</td>
<td>Thr</td>
<td>Threonine(^8)</td>
<td>ACA,ACC,ACG,ACT</td>
</tr>
<tr>
<td>18</td>
<td>V</td>
<td>Val</td>
<td>Valine(^9)</td>
<td>GTA,GTC,GTG,GTG</td>
</tr>
<tr>
<td>19</td>
<td>W</td>
<td>Trp</td>
<td>Tryptophan(^10)</td>
<td>TGG</td>
</tr>
<tr>
<td>20</td>
<td>Y</td>
<td>Tyr</td>
<td>Tyrosine(^11)</td>
<td>TAC,TAT</td>
</tr>
</tbody>
</table>

The mapping from amino acids to codons is many-to-one

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Top strand; millions of bases A,C,T,G

So we can locate exons using STFT

How to choose window size? Usual time-frequency resolution tradeoff

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Filtering interpretation

Take any base, say G:

\[ x_G(n) = 1  1  0  0  0  0  0  1  0  0  0  0  0  1  1  0  1  1  0  0  0  0  1  1  0  1  1  0  0  1  1  0 \]

\[ N \]

\[ w(n) \]

Sliding window

\[ y_G(n) \]

filter with impulse response \( h(n) \)

\[ 2\pi/3 \]

\[ 2\pi \]

Frequency response magnitude

corresponds to 13 dB

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Spectrum at $2\sqrt{3}$ as a function of base location

Gene F56F11.4 in the C-elegans chromosome III

Return to the filtering interpretation

How about designing filters to improve time-frequency resolution?

Interesting DSP problem!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Antinotch

Notch

Define two filters:

\[
\begin{bmatrix}
G(z) \\
H(z)
\end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ A(z) \end{bmatrix}
\]

\text{Allpass: } A(z) = \frac{R^2 - 2R \cos \theta z^{-1} + z^{-2}}{1 - 2R \cos \theta z^{-1} + R^2 z^{-2}}

Multistage filter design method
like the IFIR method (Neuvo, et. al, 1983)

Gene F56F11.4 in the C-elegans chromosome III

Hidden Markov models have been very successful in computational gene finding.

*Will return to it later.*
Outline

- Molecular biology background
- Computational gene-finding
- Spectral analysis (Fourier, wavelet, correlations)
  - Hidden Markov Models and sequence analysis
- New world of non-coding genes
- References

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Proteins are sequences made of 20 kinds of amino acids:
ACDEFGHIKLMNPQRSTVWY

Each amino acid is associated with a unique number called the **EIIP**: 
*Electron-ion interaction potential*

---

Given an amino acid sequence: AACDEQRIKLYXTSVDC ……

We can readily turn it into a numerical sequence $x(n)$. 

*The Fourier transform of $x(n)$ has interesting properties*

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Proteins belonging to the same functional group have something common in their Fourier transform!

Example: Fibroblast growth factor proteins

Basic bovine
PALPEDGGSGAFPPGHFKDPKRLYCKNGGF
FLRIHPDGRVDGVREKSDPHIKLQLQAEER
GVVSIKGVCNRYLAMKEDGRLLASKCVD
ECFFFERLESNNTYRSRKYSWWYVALKR
TGYKLGPKTGPQKAILFLPLMSAKS

Acidic bovine
FNLPGLNYYKKPKLLYCNGGGYFLRIILPDGT
VDGTKDRSDQHIQLQLCAEISEGEVYIKSTE
TGQFLAMDGTGLLYGSQTPNEECLFLERLE
ENHYNTYISKKHAEKHWFWGLKKNRSKLG
PRTHFGQKAILFLPLPLPVSSD
Let $x(n)$ and $y(n)$ be proteins which have a function in common. Then the product of Fourier transforms exhibits a sharp isolated peak!

Proteins work by recognizing other molecules from spatial periodic components!

**Resonant recognition model (RRM), Cosic, 1994.**

Lots of good physics behind this. See references in Cosic, 1994.
some other molecule, e.g., ligand

protein molecule

Fits like a puzzle piece. That’s how beautifully enzymes work!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Protein group: hemoglobins

Hemoglobins are oxygen carriers in the red blood cells.

Glucagons are proteins (peptide hormones) which affect glucose level in blood. Made by alpha-cells in pancreas.
Examples of other functional groups of proteins.


<table>
<thead>
<tr>
<th>PROTEIN SEQUENCES</th>
<th>Frequency</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>oncogenes</td>
<td>0.03130</td>
<td>46</td>
</tr>
<tr>
<td>kinases</td>
<td>0.42969</td>
<td>8</td>
</tr>
<tr>
<td>fibrinogens</td>
<td>0.44230</td>
<td>5</td>
</tr>
<tr>
<td>ACII receptors</td>
<td>0.49219</td>
<td>21</td>
</tr>
<tr>
<td>phages' repressors</td>
<td>0.10547</td>
<td>4</td>
</tr>
<tr>
<td>bacterial repress.</td>
<td>0.08398</td>
<td>4</td>
</tr>
<tr>
<td>heat shock proteins</td>
<td>0.09473</td>
<td>10</td>
</tr>
<tr>
<td>interferons</td>
<td>0.08203</td>
<td>18</td>
</tr>
<tr>
<td>hemoglobins</td>
<td>0.02340</td>
<td>187</td>
</tr>
<tr>
<td>signal proteins</td>
<td>0.14063</td>
<td>5</td>
</tr>
<tr>
<td>protease inhibitors</td>
<td>0.35550</td>
<td>27</td>
</tr>
<tr>
<td>proteases</td>
<td>0.37700</td>
<td>80</td>
</tr>
<tr>
<td>restriction enzymes</td>
<td>0.29102</td>
<td>3</td>
</tr>
<tr>
<td>amylases</td>
<td>0.41211</td>
<td>12</td>
</tr>
<tr>
<td>neurotoxins</td>
<td>0.07031</td>
<td>16</td>
</tr>
<tr>
<td>growth factors</td>
<td>0.29297</td>
<td>105</td>
</tr>
<tr>
<td>ins.-like(IGF I,II)</td>
<td>0.49220</td>
<td>12</td>
</tr>
<tr>
<td>FGFs</td>
<td>0.45120</td>
<td>7</td>
</tr>
<tr>
<td>glucagonos</td>
<td>0.32030</td>
<td>13</td>
</tr>
<tr>
<td>homeo box proteins</td>
<td>0.04590</td>
<td>9</td>
</tr>
<tr>
<td>cytochromes B</td>
<td>0.05900</td>
<td>16</td>
</tr>
<tr>
<td>cytochromes C</td>
<td>0.47656</td>
<td>38</td>
</tr>
<tr>
<td>myoglobins</td>
<td>0.08200</td>
<td>49</td>
</tr>
<tr>
<td>lysozymes</td>
<td>0.32810</td>
<td>15</td>
</tr>
<tr>
<td>phospholipases</td>
<td>0.04300</td>
<td>29</td>
</tr>
<tr>
<td>actins</td>
<td>0.48000</td>
<td>12</td>
</tr>
<tr>
<td>myosins</td>
<td>0.34000</td>
<td>11</td>
</tr>
<tr>
<td>RNA polymerases</td>
<td>0.35693</td>
<td>10</td>
</tr>
</tbody>
</table>

Frequency normalized so that $2^0$ corresponds to 1.
By **localizing** the spatial domain region which has the greatest influence at the **resonance** frequency, one can identify the small **region** in a large protein molecule which is **responsible** for a particular function.

**Hot spots of the protein**

- Usual tradeoff between frequency localization and time localization.
- **Wavelet transform**: natural candidate for this.

Long-range correlation in DNA sequences

DNA  AACTGGCATCCGGGAATAAGGTC

\( x_A(n) \)  1 1 0 0 0 0 10 0 0 0 0 110 1 1 0 0 00

\( r_A(k) \)

Decays very slowly!

Lot of correlation between bases millions away

Long-range correlation or 1/f property

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Fourier transform pair: $\frac{1}{|f|^\alpha} \iff c \cdot |t|^\alpha - 1$

called 1/f property for any $\alpha > 0$.

1/f behavior is equivalent to long range correlation in time.

Examples:

- $\alpha = 1$ for traditional 1/f noise.
- $\alpha = 2$ for Brownian noise.

---

*Papoulis, Systems and transforms, 1968*

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
History of 1/f behavior in DNA

**Peng**, et al., Nature, March 92 (studied genes with introns).

**Voss**, Physical review letters, June 92 (studied human DNA, other organisms).


**Hausdroff** and **Peng**, Physical review E, Aug. 96 (multiscale randomness).

**Early work on theory:**

1/f behavior is well known in the physical world: Noise in resistors, sunspot activity, flood levels, audio spectra, all exhibit 1/f feature.

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Example: Bacteria aquifex aeolicus, size 1.55 Mb.

This is a typical DNA power spectrum

PSD for base A; 1 million bases used

PSD of base A in randomly generated “DNA”.

No evidence of any 1/f behavior

Why is there long range correlation in DNA?

If all life evolved from a common ancestor, then today’s long DNA must have evolved from short DNAs of early life.

DNA size evolution

- Earliest life: few 1000 bases (half a billion years ago)
- Today’s smallest bacteria: few million bases
- Mammals like us: few billion bases.

Evolution model: duplicate and mutate model.

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Suppose we generate a long binary sequence $x(n)$ as follows:

- Start from a short binary seed $s(n)$.
- Duplicate and mutate randomly with small error probability $p$.
- Concatenate the result to $s(n)$.
- Keep repeating this to get the long sequence $x(n)$.

Can you prove that $x(n)$ has the $1/f$ property?

*W. Li, Physical review A, American Physical Society, May 1991*
End of this part
Outline

• Molecular biology background

• Computational gene-finding

• Spectral analysis (Fourier, wavelet, correlations)

• Hidden Markov Models and sequence analysis

• New world of non-coding genes

• References
Markov models

**DNA sequence:** AACTGAGGTACAAATTTCGATCTC

![Diagram of Markov model]

State transition matrix

\[
\begin{pmatrix}
A & C & T & G \\
A & 0.1 & 0.2 & 0.4 & 0.3 \\
C & 0.2 & 0.5 & 0.1 & 0.2 \\
T & 0.5 & 0.2 & 0.1 & 0.2 \\
G & 0.3 & 0.1 & 0.4 & 0.2
\end{pmatrix}
\]

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Application of Markov models

Given a DNA sequence: \( X = x(1) x(2) x(3) \ldots \ x(N) \)

And given a Markov model \( \square \), we can calculate:

\[
P(X) = P(x(1)) \times P(x(1) \text{ to } x(2)) \times P(x(2) \text{ to } x(3)) \times \ldots
\]

**Probability that sequence \( X \) is generated by model \( \square \):**

Given a set of models:

\[
\begin{align*}
\square_1 & \quad \text{Model 1} \\
\square_2 & \quad \text{Model 2} \\
\vdots & \quad \vdots \\
\square_K & \quad \text{Model K}
\end{align*}
\]

we can find the model which most likely generated the sequence \( X \).

The models are obtained by **training** with known sequences.

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Hidden Markov Models (HMM)

*In an HMM, states are not the same thing as outputs.*
Example! States: 1, 2, 3 Outputs: A, C, T, G

States could be *exon, intron, Cpg island, etc.* Outputs could be *bases.*

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
HMM example:

State transition matrix:

\[
\begin{pmatrix}
1 & 2 & 3 \\
1 & 0.3 & 0.7 & 0.0 \\
2 & 0.0 & 0.4 & 0.6 \\
3 & 0.9 & 0.0 & 0.1 \\
\end{pmatrix}
\]

Output matrix:

\[
\begin{pmatrix}
A & C & T & G \\
1 & 0.3 & 0.1 & 0.4 & 0.2 \\
2 & 0.5 & 0.3 & 0.1 & 0.1 \\
3 & 0.1 & 0.3 & 0.4 & 0.2 \\
\end{pmatrix}
\]

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
HMM was used in speech recognition in 80’s (Rabiner).

The bioinformatics community learnt the basic ideas from Larry Rabiner’s famous IEEE tutorial (Proc. of the IEEE, 1989)

Today HMM is routinely used in genomics and proteomics:

• Gene identification
• DNA sequence alignment (big area; lots of problems)
• Identification of CpG islands in DNA


P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
HMM is a finite state machine (FSM) and represents **regular grammars**.

**Regular grammar**

Only production-rules of the form: \( W \rightarrow aW \)

\( W: \) nonterminal \hspace{1cm} \( a: \) terminal

*Example: suppose the grammar is defined by these rules:*

\[
W \rightarrow AW \hspace{1cm} W \rightarrow TW \hspace{1cm} W \rightarrow CW
\]

*Example of a string generated by this grammar:*

\[
W \rightarrow AW \rightarrow AAW \rightarrow AACW \rightarrow AACTW \rightarrow AACT
\]

**Theorem:** HMM is equivalent to **stochastic** regular grammars

**Stochastic means:** each rule is used with a certain probability

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Regular grammar example:
\[ W \rightarrow AW \rightarrow AAW \rightarrow AACW \rightarrow AACTW \rightarrow AACT \]

Context free grammar (CFG):
Production rules of the form: \[ W \rightarrow \square \]
\( W \): nonterminal \quad \square: \) string of terminal and or nonterminals

Example: grammar with production rules:
\[ W \rightarrow AWA \quad W \rightarrow CWC \quad W \rightarrow TWT \quad W \rightarrow GWG \quad W \rightarrow \text{null} \]

Example of sequence generated:
\[ W \rightarrow AWA \rightarrow ATWTA \rightarrow ATCWCTA \rightarrow ATCCTA \]
This is a symmetric sequence (palindrome)

Grammar which generates precisely the set of all palindromes cannot be regular; it has to be a context free grammar.

Stochastic context free grammar (SCFG): the rules are used stochastically.

The palindrome language cannot be generated by HMM. We need SCFG for that.
Chomsky’s hierarchy of grammars (1956)

unrestricted
context sensitive
context free
regular

SCFG
non-coding genes
ncRNAs
siRNAs

these have palindrome components

HMM
introns
exons
CpG islands
intergenic

Noan Chomsky, 1928-- computational linguist, MIT

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Outline

• Molecular biology background

• Computational gene-finding

• Spectral analysis (Fourier, wavelet, correlations)

• Hidden Markov Models and sequence analysis

• New awareness of non-coding genes

• References

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Recent discovery:

Intergenic space has lots and lots of genes! Not junk after all.

But these are different kinds of genes. They generate RNA which do not code for proteins.

**RNA-genes or noncoding genes.**

Noncoding RNA (ncRNA)

The RNA remains in the cell and performs its own functions!

*W. W. Gibbs, The unseen genome, Scientific American, 11/03*

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Recall Crick’s Central dogma of molecular biology:

\[
\text{DNA} \xrightarrow{\text{transcript}} \text{mRNA} \xrightarrow{\text{translate}} \text{protein}
\]

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
RNA molecules acknowledged by central dogma

**mRNA: messenger RNA**
The gene is transcribed into mRNA which carries the genetic code to ribosome

**tRNA!transfer RNA**
helps in translation of mRNA to protein

**rRNA: ribosomal RNA!**
helps in translation of mRNA to protein

*A few others like snoRNA, etc. These are the classic non-coding RNAs.*

**But now biologists have found many more ncRNAs.**
**Central dogma of molecular biology challenged!**

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
The heroic detective story

There was once a C. Elegans baby that would not grow up beyond the first (of four) larval stage; kept repeating stage 1. 
Getting bigger but not growing up.

John Travis, “Biological dark matter”, Science News, 1/02

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004,Vancouver
There was a **defective gene** responsible for this.

In the healthy worm the gene’s function was to release a **tiny RNA molecule** (22 bases long) into the cell.

This RNA had its own function: **regulate** other protein coding genes responsible for normal growth.

In the **defective worm** the gene was not generating this RNA properly. This was the first nc-RNA to be taken seriously (other than the classic ones).

*Ambrose et al., 1993 (Dartmouth medical school, Hanover, N. H).*

*P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver*
Today nc-RNA genes are recognized to be extremely crucial to the functioning of cells.

**Hereditary information is carried by**

1. Protein-coding genes (known for many years).
2. ncRNA genes.
3. Epigenetic layers
What is there in it for the signal processor?

We know protein coding genes can be identified on the computer.

ncRNA genes are much more difficult to identify on the computer.

Still an open problem in computational molecular biology! But why is it so challenging?

- ncRNA could be very small (e.g., 22 bases)
- There is no codon bias (period 3) or open reading frame (ORF)
- No start and stop codons
- Cannot go by size. Protein coding genes with 7 bases are known!
- Other reason: we have to examine secondary structure (see later).

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Computational identification of ncRNA genes

A new discipline called comparative genome analysis helps to distinguish coding genes from nc-genes.

Does not work perfectly yet

Example 1
360-base bacterial regulatory ncRNA CsrB gene: (first thought to be protein coding gene)

Example 2
The plant (Medicago) ENOD40 gene was thought to be an ncRNA gene based on sequence analysis. Recently based on comparative genome analysis, found to encode two tiny proteins (13 and 27 amino acids long).

S. R. Eddy, Nature reviews, GENETICS, 12/01

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Comparative genomics

If two or more species have a common stretch of DNA then it is probably doing something important. Otherwise nature would not have conserved it for millions of years.

To compare genomes, one has to solve the alignment problem.

\[
\begin{align*}
\text{xxAATAGCGA} & \ldots \text{AATAC} \ldots \text{AAATACCG} \\
\text{xxxxxAATAGCGA} & \ldots \text{AATAC} \ldots \text{AAATACCG} \\
\text{xxxxxAAGAGCGA} & \ldots \text{AATAC} \ldots \text{AAAAGTGCCG} \\
\text{xxxxxAAAGCGA} & \ldots \text{AATAC} \ldots \text{AAATAAACCG}
\end{align*}
\]

Multiple-alignment problem with gaps and mutations

Scoring problem

Hidden Markov models, again useful.

Lots of good problems for theoreticians!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
The human genome has been compared with:

Cows
Dogs
Pigs
Rats
7 others …

And there were 1,200 common segments; 154 in intergenic area.

Study by NHGRI (National Human Genome research institute)

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Examples

- Many nc-RNA genes have been found in flies, worms, humans.
- E. Coli bacterium has 4200 protein coding genes and several hundreded nc-RNA genes.
- About 50% of genes in mice could be nc-RNA genes.
- C. Elegans probably has over 200 micro-RNA genes (20%).

Intergenic space = biological dark matter?

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Number of protein-coding genes does not scale well with organism’s complexity

- Worms have only twice as many protein-coding genes as bacteria
- Humans: probably only twice as much (about 27,000)
- Rice plant: more genes than humans!

But apparently the number of ncRNA genes does!

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Functionality of ncRNAs depends mostly on their **secondary structure**.

Notice the secondary structure created by base pairing in blue shaded areas

**dsrA RNA in E. Coli**

See S. R. Eddy, *Nature Reviews*, 12/01 for many examples and detailed discussions

tRNA molecule (clover-leaf form)

http://www.ebi.ac.uk/microarray/biology_intro_files/tRNA.htm

Notice amazing amount of secondary structure
Linear sequence representing an ncRNA-gene

Compuational biologists try to identify ncRNA genes by looking for the palindrome patterns buried in the linear sequence.

HMMs cannot represent palindromes!

We need context-free grammars, and the search is more difficult.

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
These two sequences will probably fold into the same secondary structure or shape. And that is what really matters as far as biochemical function is concerned.

Finding a particular ncRNA gene does not necessarily mean looking for a particular sequence. We really are looking for hidden palindromes at appropriate places.
Routine steps in the application of HMM

Given the HMM and an output sequence $y(1), y(2), \ldots$ how to compute the state sequence which most likely generated it?

*Viterbi’s algorithm* (same as the one in digital communications)

Given the HMM and an output sequence $y(1), y(2), \ldots$ how to compute the probability that the HMM generates this?

*Forward-backward algorithm*

How to adjust the model parameters $S$ and $P$ such that they are optimal for an application, e.g., to represent exons?

*Training; Expectation Maximization algorithm (Baum-Welch).*
Folded RNA sequence

HMMs cannot represent palindromes!

**We need context-free grammars**

How to systematically develop algorithms based on such grammars?

**For example**

- Is there a Viterbi-like algorithm?
- Is there a forward-backward algorithm?
- Is there a Expectation-Maximization-like algorithm?

*Need FAST algorithms because genomes are looong!*

**Ongoing research topic in computational molecular biology today.**

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Biology today is not just wet stuff in smelly labs!

Molecular biology involves signal processing, computer science, mathematics, informatics, all coming together wonderfully!

End of this part

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Outline

• Molecular biology background
• Computational gene-finding
• Spectral analysis (Fourier, wavelet, correlations)
• Hidden Markov Models and sequence analysis
• New world of non-coding genes
• References

P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
REFERENCES FOR THE GENOMIC SIGNAL PROCESSING TALK

Plenary lecture by Prof. P. P. Vaidyanathan, Caltech, Pasadena, CA
“Genomic signal processing”, ISCAS-2004 Vancouver, Canada, May 2004

http://www.systems.caltech.edu/dsp/IsicasGenomeTalkRef/

I have tried to categorize the papers into subtopics but this has been difficult. Many papers can easily belong in more than one category. So please do not overlook any of these. The selection here is by no means extensive. It is based entirely on my personal taste. Perhaps a good list to start with, to teach from, and so forth — P.P.V.
The great paper
The paper which started it all ...


Books and Tutorials


P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004, Vancouver
Signal-processing flavor (DNA/Protein)


Gene prediction


Long range correlations, $1/f$ behavior, statistics


Noncoding RNA, Noncoding genes


DNA microarrays


P. P. Vaidyanathan, ISCAS Plenary, 5/24/2004,Vancouver