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:**

...ATTCGAAGATTTCAACGGGAAAA...

DNA

2. Sequences from an alphabet of **size twenty:**

AACWYDEFGHIKLMNPQRSTVAPPQR

Protein

Size-4 alphabet:

A, C, T, G: bases (also called or nucleotides)

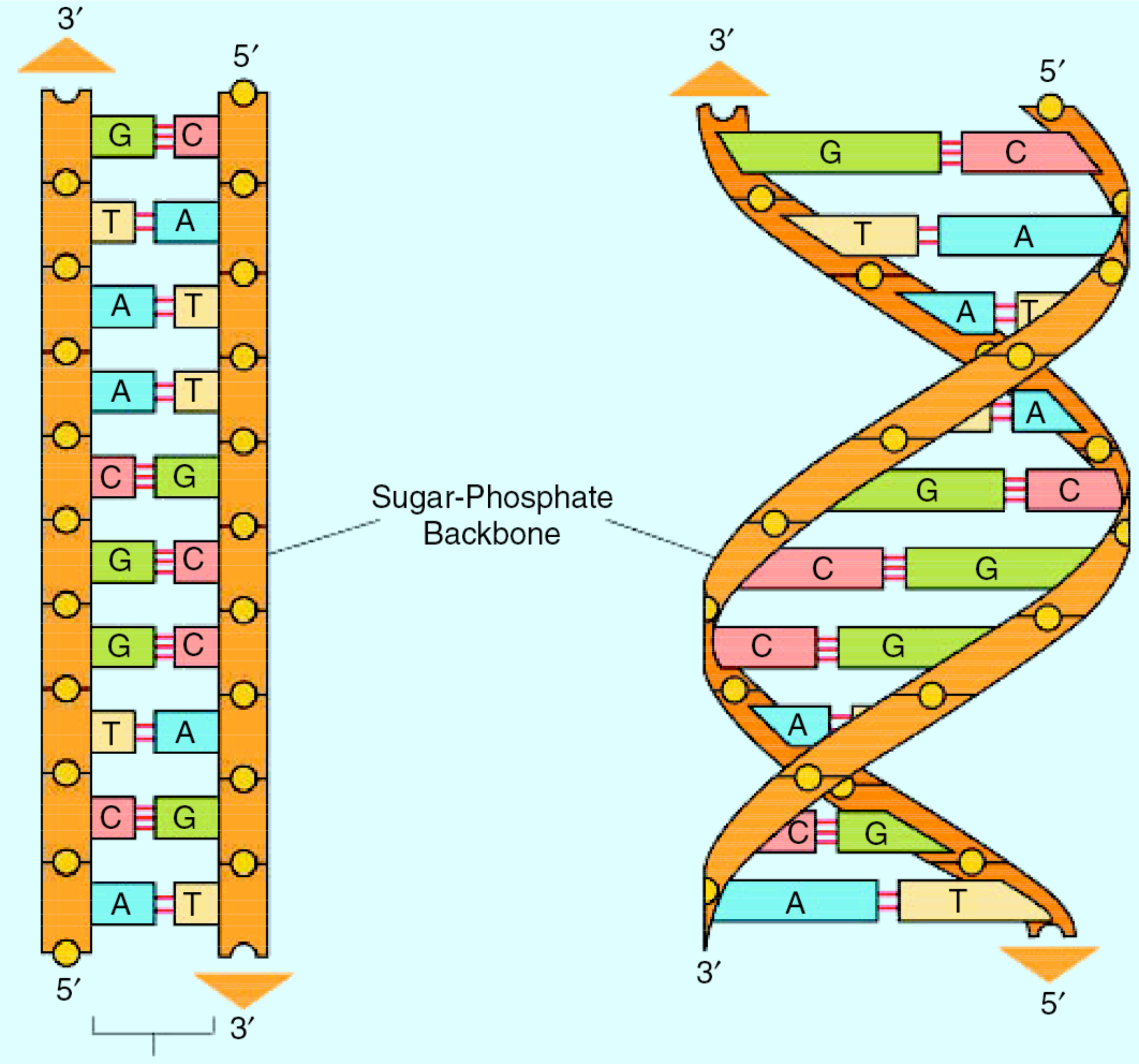
DNA sequences (genomes) are made of these.

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

Adenine **T**hymine **C**ytosine **G**uanine

or Uracil (in RNA)

DNA: deoxyribonucleic acid
RNA:ribonucleic acid



Hydrogen bond

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

Alberts, Bray, Johnson, Lewis, Raff, Roberts, and Walter

DNA molecule in the living cell (usually in nucleus)

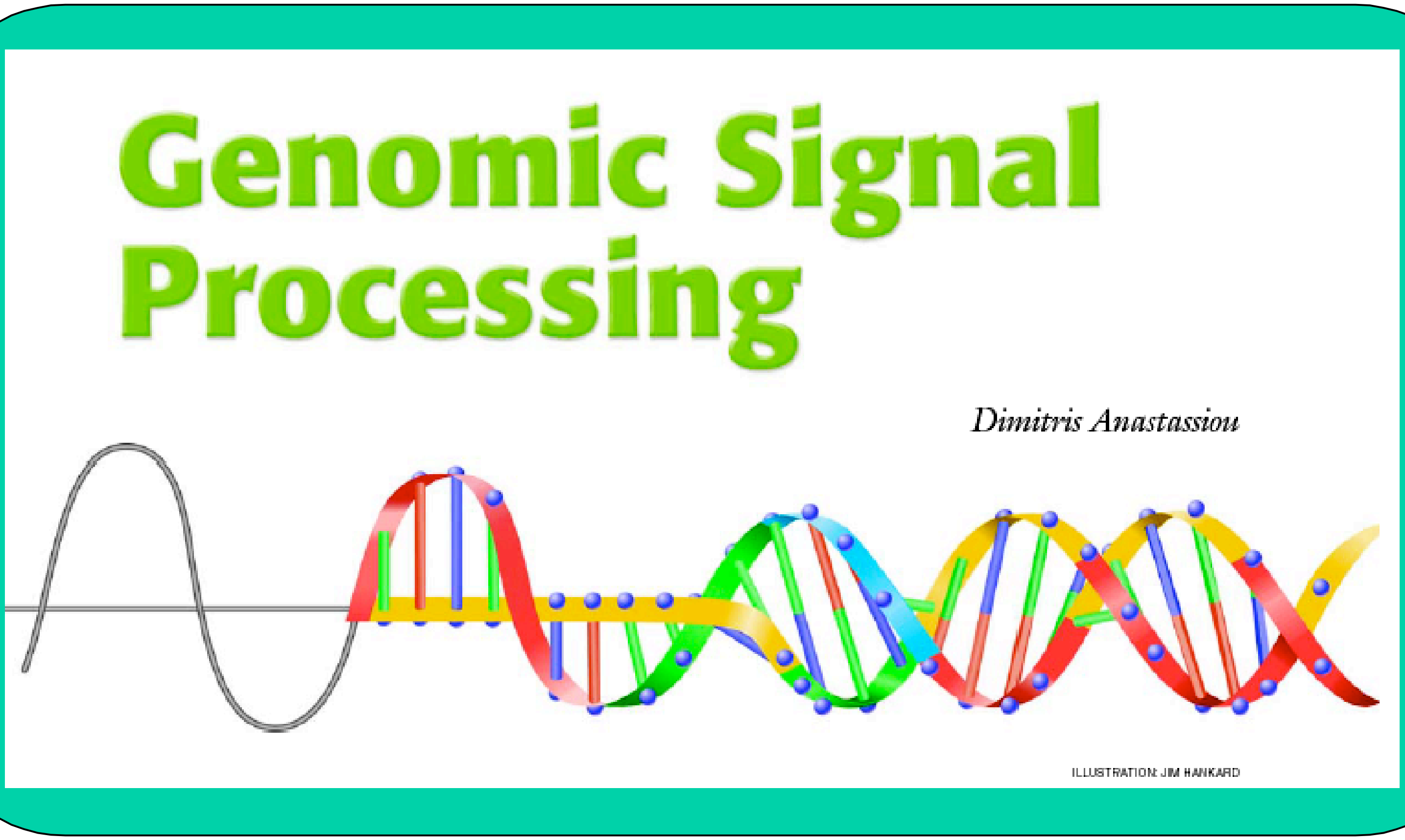
Complementary Strands in the Double Helix

A = T
C ≡ G

Great place to get started, and a great reference

A good introductory article (signal processing aspects)

Dimitris Anastassiou, IEEE Signal Processing Magazine, July 2001



Genomic Signal Processing

Dimitris Anastassiou

ILLUSTRATION: JIM HANKARD

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

The twenty natural amino acids

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

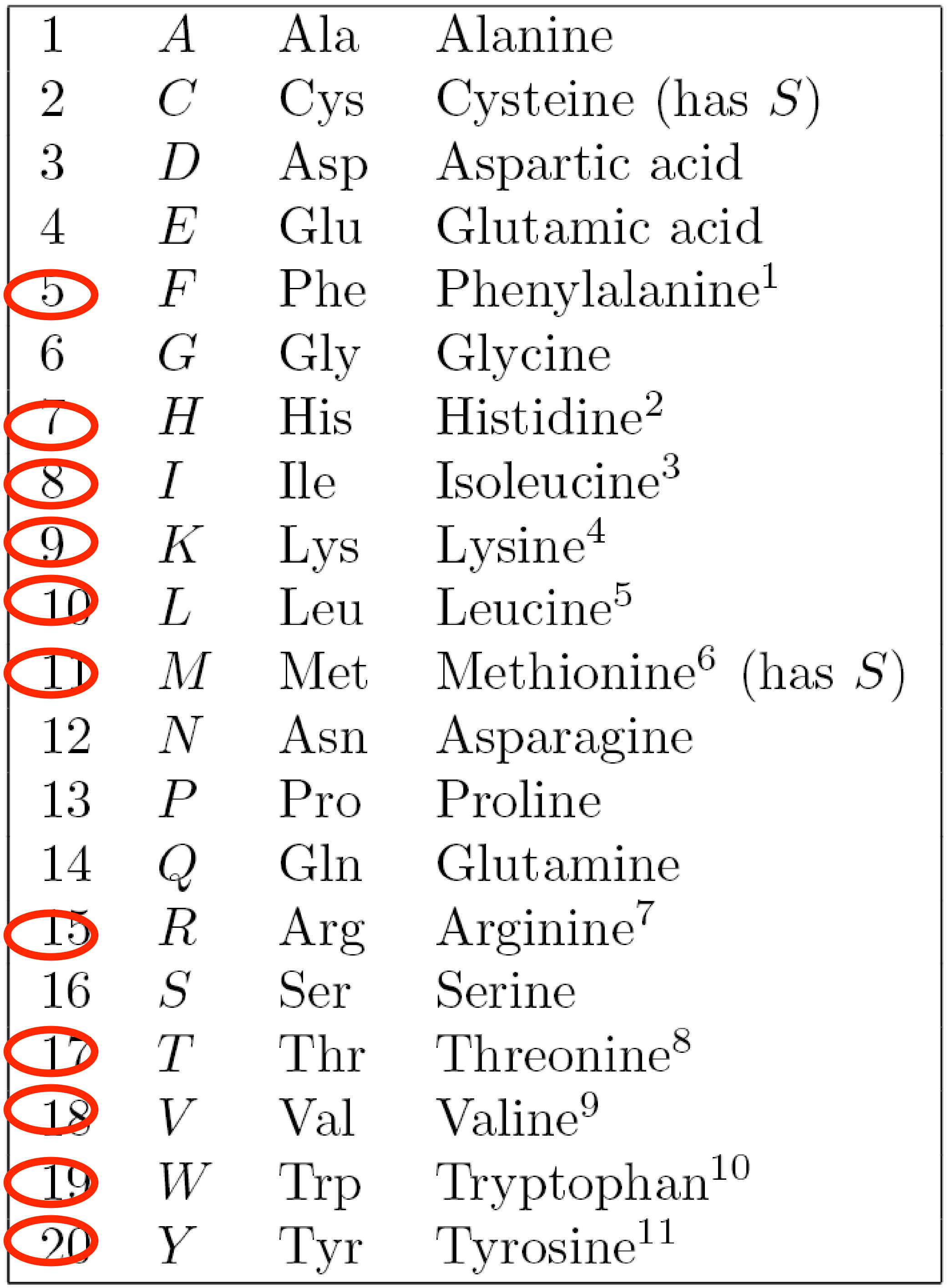
11 essential amino acids.

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.



1	A	Ala	Alanine
2	C	Cys	Cysteine (has S)
3	D	Asp	Aspartic acid
4	E	Glu	Glutamic acid
5	F	Phe	Phenylalanine[1]
6	G	Gly	Glycine
7	H	His	Histidine[2]
8	I	Ile	Isoleucine[3]
9	K	Lys	Lysine[4]
10	L	Leu	Leucine[5]
11	M	Met	Methionine[6] (has S)
12	N	Asn	Asparagine
13	P	Pro	Proline
14	Q	Gln	Glutamine
15	R	Arg	Arginine[7]
16	S	Ser	Serine
17	T	Thr	Threonine[8]
18	V	Val	Valine[9]
19	W	Trp	Tryptophan[10]
20	Y	Tyr	Tyrosine[11]

Protein Example

Fibroblast growth factor proteins

Basic bovine

PALPEDGGSGAFPPGHFKDPKRLYCKNGGF
FLRIHPDGRVDGVREKSDPHIKLQLQAEER
GVVSIKGVCANRYLAMKEDGRLLASKCVTD
ECFFFERLESNNYNTYRSRKYSSWYVALKR
TGQYKLGPKTGPGQKAILFLPMSAKS

length 146

Acidic bovine

FNLPLGNYKKPKLLYCSNGGYFLRILPDGT
VDGTKDRSDQHIQLQLCAESIGEVYIKSTE
TGQFLAMDTDGLLYGSQTPNEECLFLERLE
ENHYNTYISKKHAEKHWFVGLKKNGRSKLG
PRTHFGQKAILFLPLPVSSD

length 140

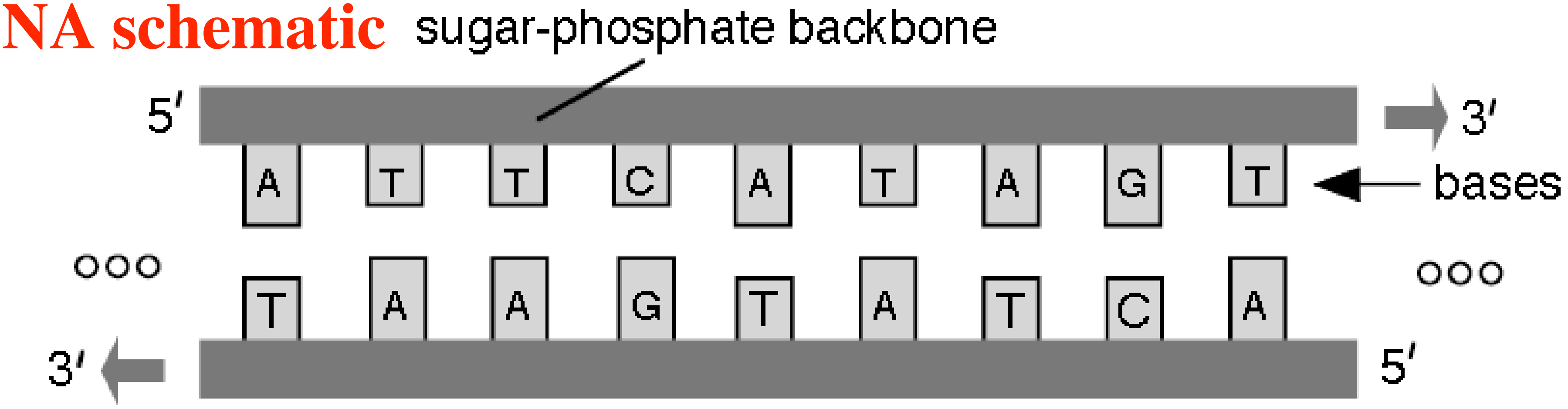
Will return to these and talk about their Fourier transforms

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.

DNA schematic



Bacterial DNA: few **million** bases;   Human DNA: three **billion** bases

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.

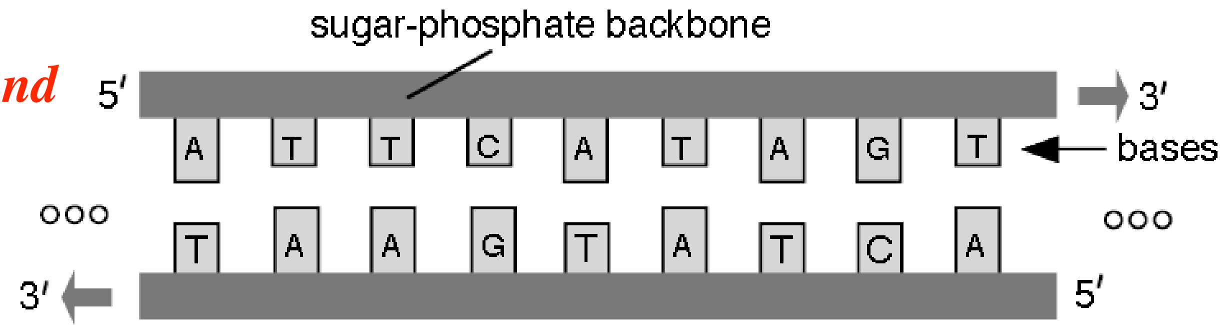
Actual physical size:

- human DNA in any cell stretches out to **2 yards.**
- DNA in all 5 trillion cells in humans:

earth
ACTTAAGGCCAAAGATCAGG ...
sun

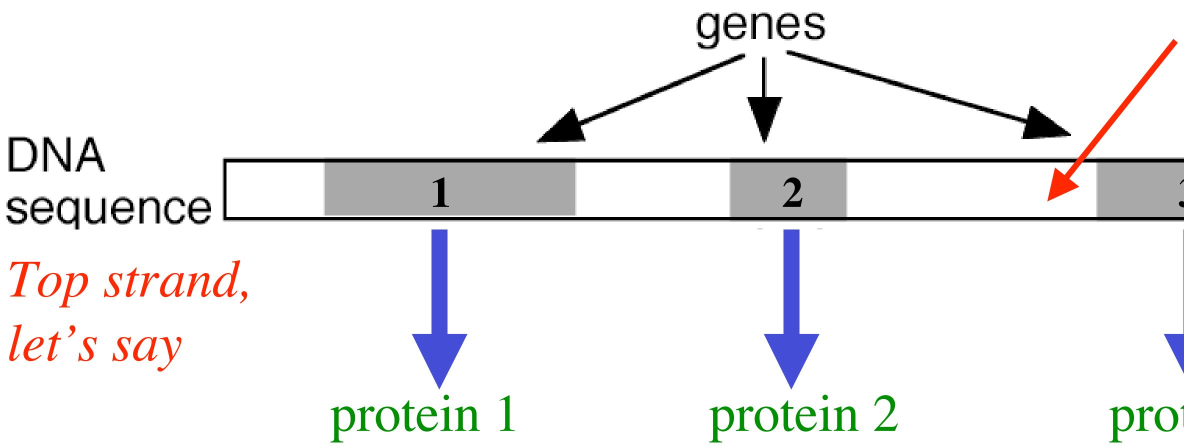
Covers it 50 times over

What do genes do?



top strand

bottom strand



Top strand, let's say

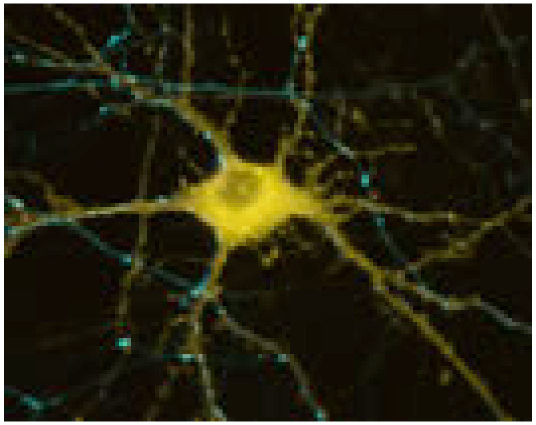
Intergenic spaces; contain A,T,C,G too!

Lots of protein in the cell, inside and outside nucleus

All cells in a given organism have the same DNA; same set of genes.

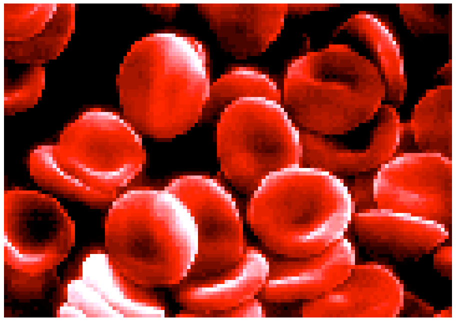
But **different genes are expressed(i.e., functional)** in different cells.

That's why **brain cells** look different from **blood cells,** and so forth.



Brain cell

http://www-biology.ucsd.edu/news/article_112901.html



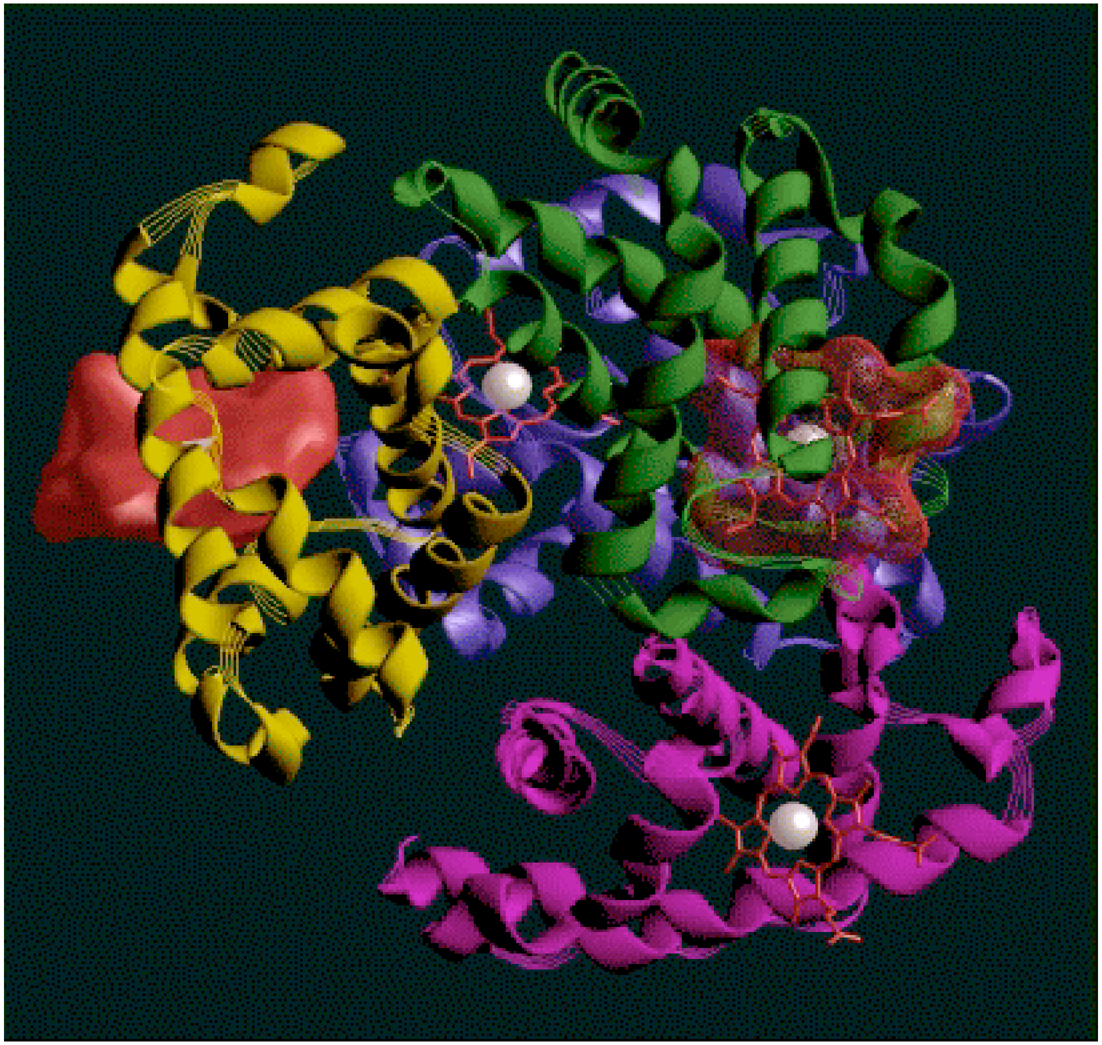
Red blood cells

http://www.cellsalive.com/gallery.htm

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

Each gene makes a different protein.

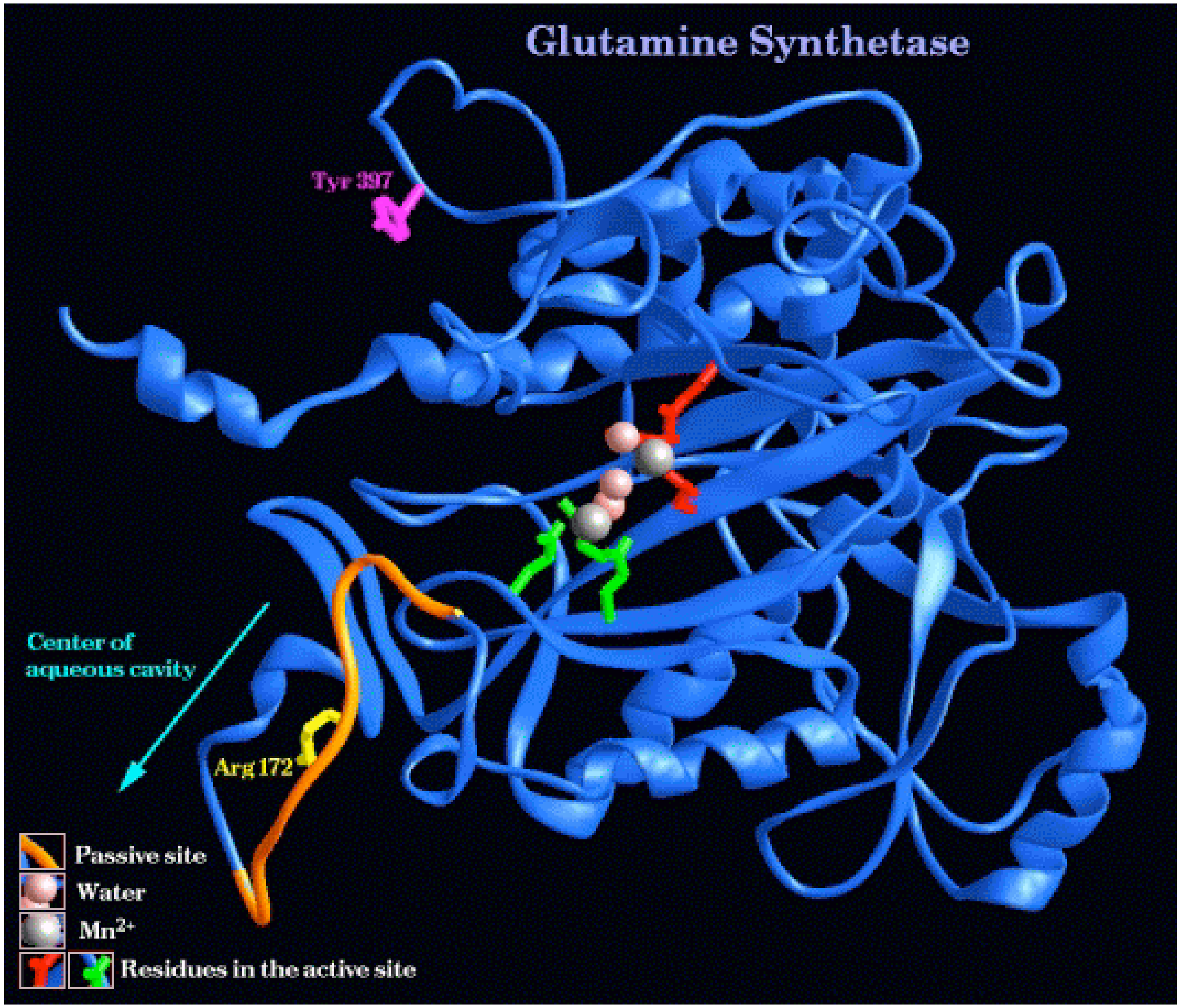
Example of a Protein: Hemoglobin (oxy, human)



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

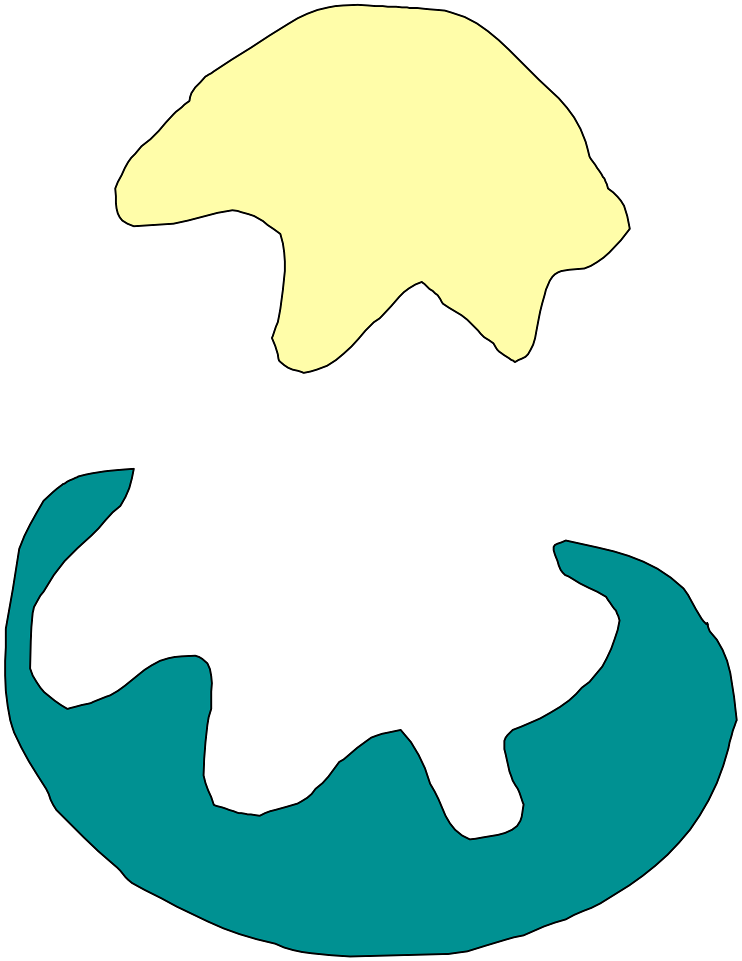
Sequence of amino acids. Folds into beautiful 3D shapes. Necessary for function.

Example of a protein (an enzyme)

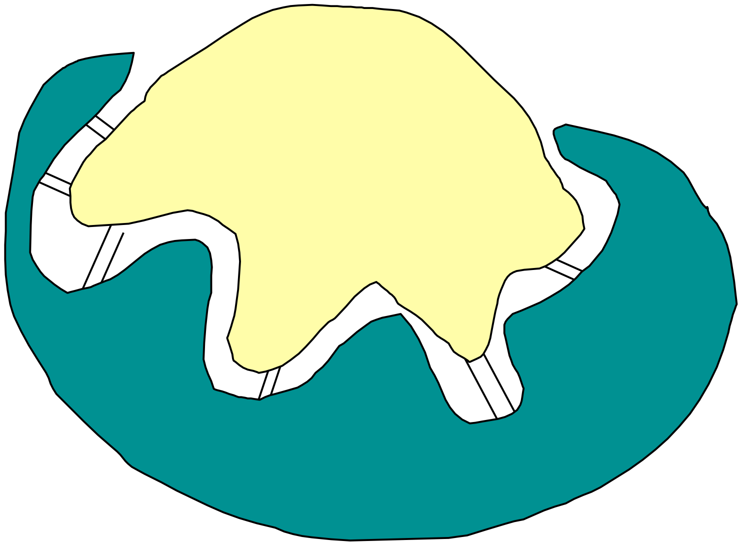


http://www.biochem.szote.u-szeged.hu/astrojan/protein2.htm

some other molecule,
e.g., ligand

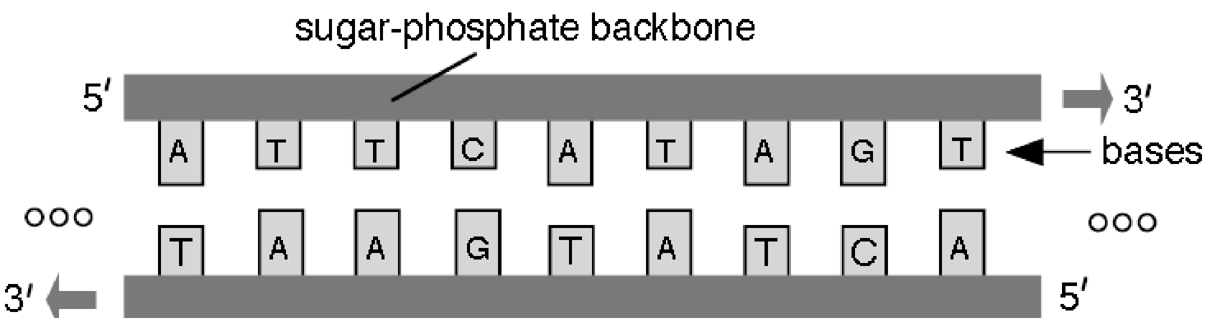


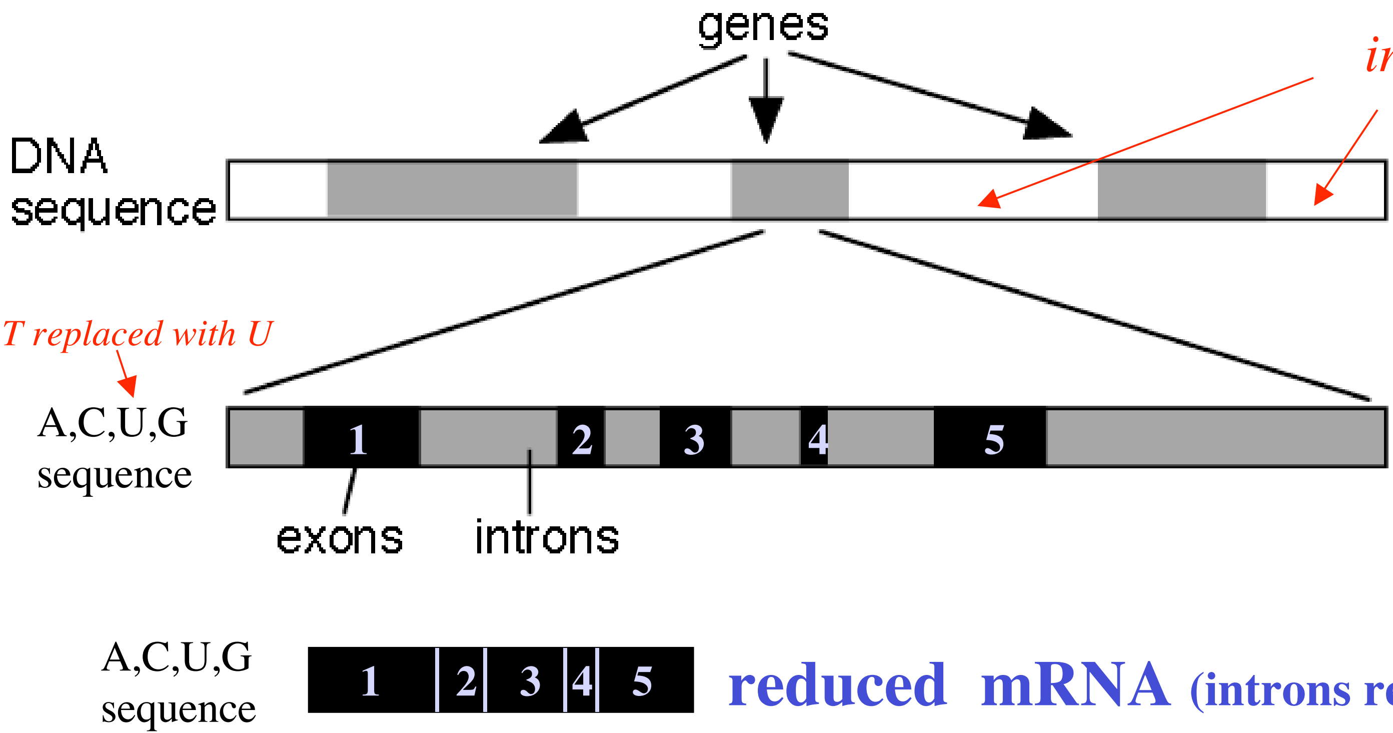
protein molecule



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

Generation of a protein from a gene





intergenic spaces

Top strand;
millions of bases
A,C,T,G

T replaced with U

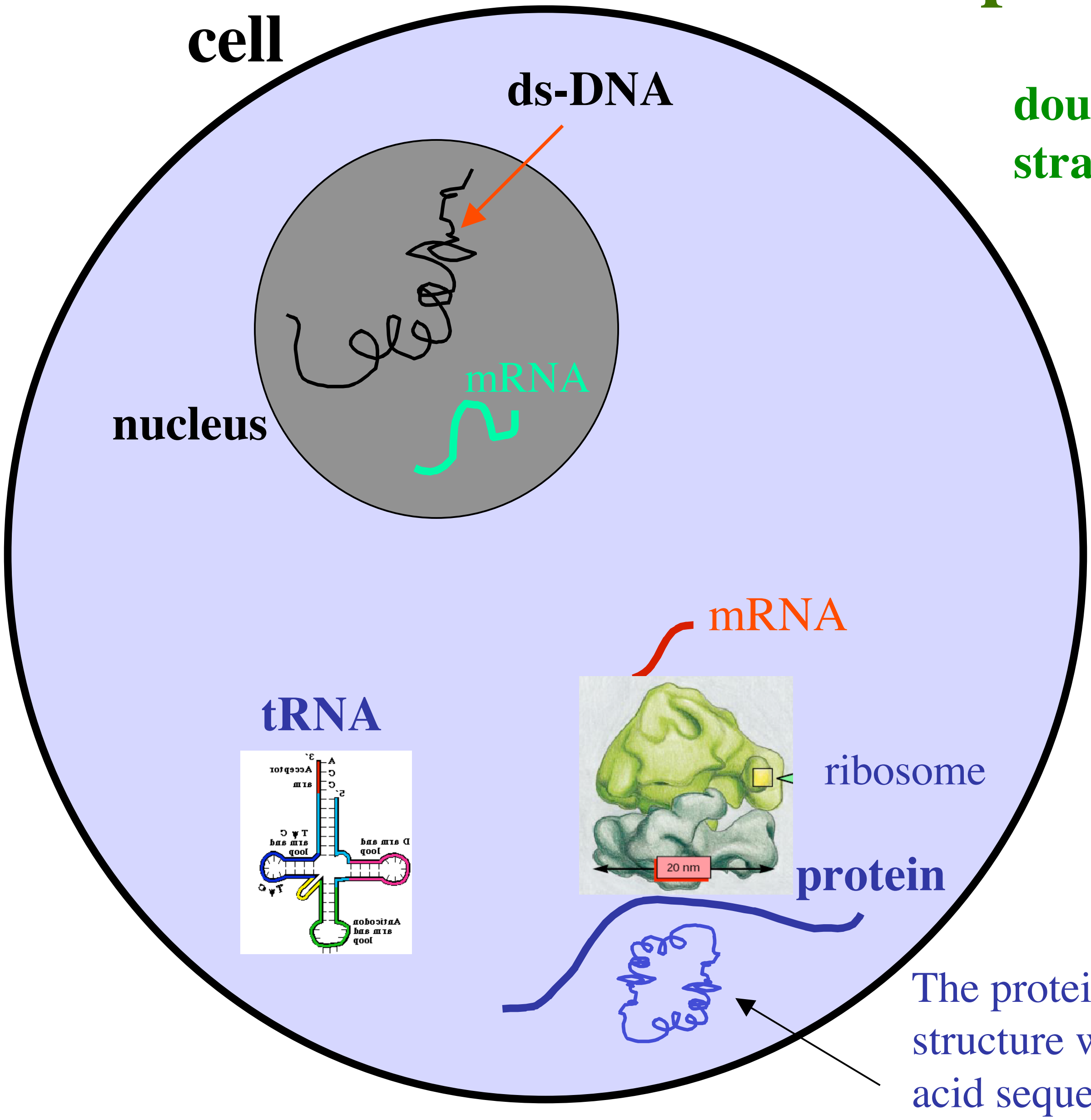
Gene copied into mRNA
(transcription)

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,Vancover

Generation of a protein from a gene



double strand opened up, one strand copied as an RNA

introns removed and mRNA reduced by splicing

ribosome converts mRNA into protein

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)

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

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

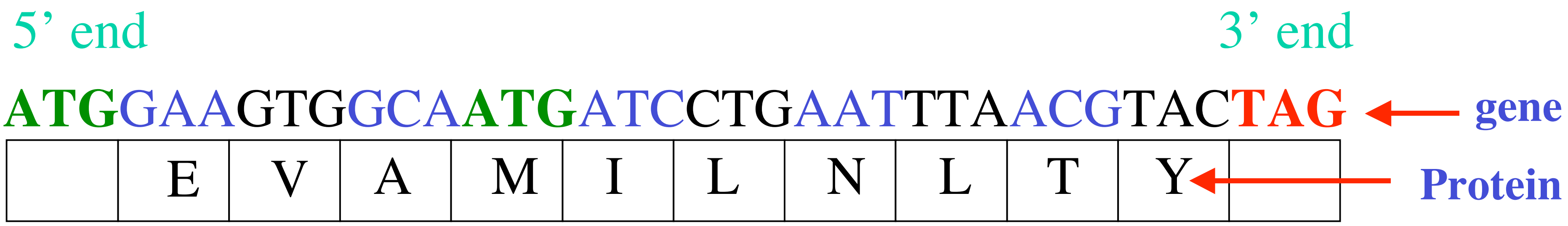
In recent years the central dogma has been challenged!

Role of codons

Gene from DNA scanned from 5’ to 3’ end:

5’ ATGGAAGTGGCAATGATCCTGAATTTAACGTACTAG 3’

The gene is interpreted in groups of three bases called **codons**.



5’ end 3’ end

ATGGAAGTGGCAATGATCCTGAATTTAACGTACTAG ← gene

	E	V	A	M	I	L	N	L	T	Y	

← Protein

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

The genetic code

codon amino acid

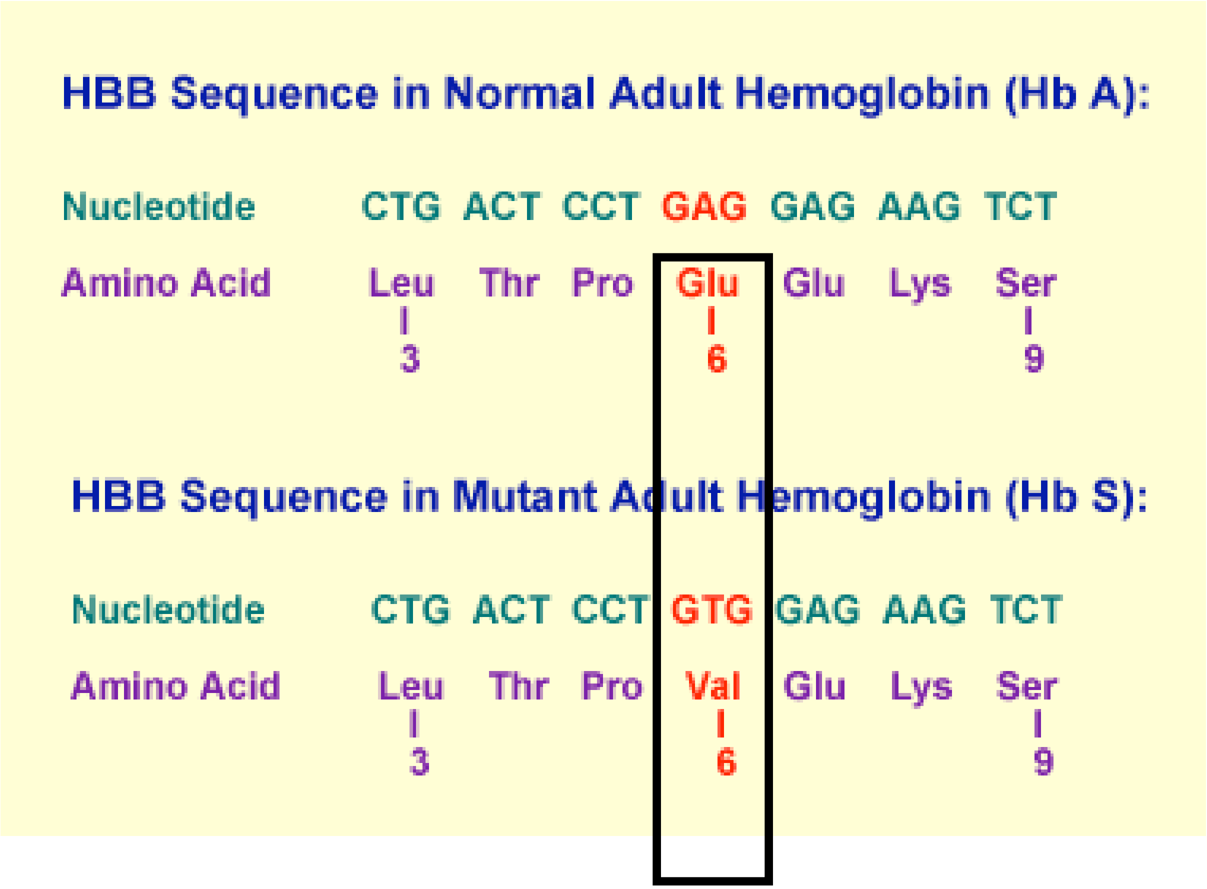
AAA: K (Lys)	GAA: E (Glu)	TAA: STOP	CAA: Q (Gln)
AAG: K (Lys)	GAG: E (Glu)	TAG: STOP	CAG: Q (Gln)
AAT: N (Asn)	GAT: D (Asp)	TAT: Y (Tyr)	CAT: H (His)
AAC: N (Asn)	GAC: D (Asp)	TAC: Y (Tyr)	CAC: H (His)
AGA: R (Arg)	GGA: G (Gly)	TGA: STOP	CGA: R (Arg)
AGG: R (Arg)	GGG: G (Gly)	TGG: W (Trp)	CGG: R (Arg)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

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.



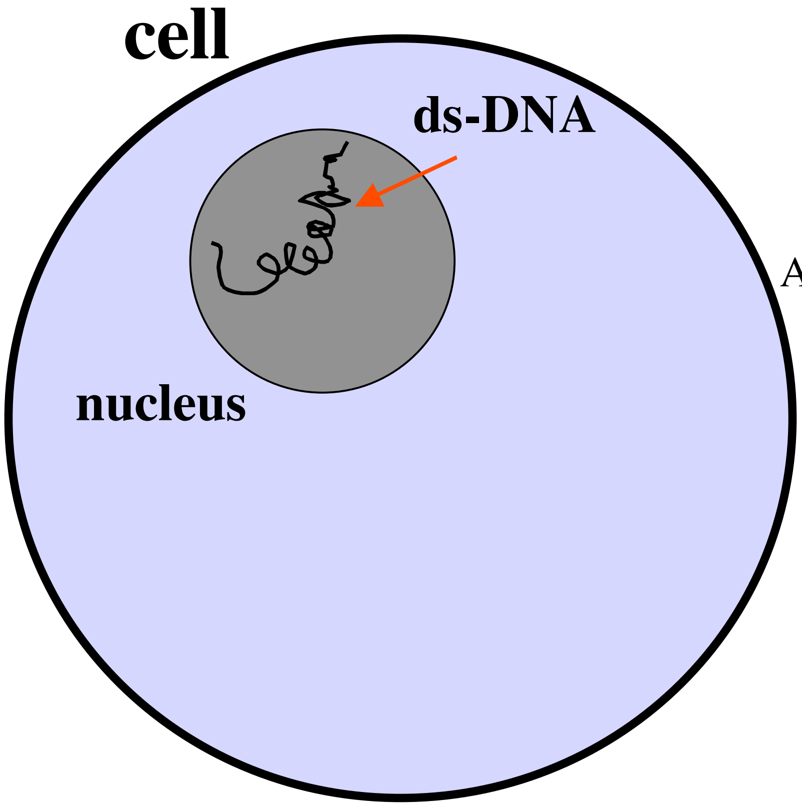
HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9

http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/hbb.shtml



DNA replicates itself when the cell divides.

AATATAGACCGACCCTAAGTAAAATAGACCTAGTAGA

1 error per billion bases.

$$\boxed{P_e = 10^{-9}}$$

Built-in proof reading system called mismatch-pair system

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 erorr per slide

Beginning of the history of molecular biology:

J. D. Watson, and F. H. C. Crick, A structure for DNA, ***Nature***, 4/1953

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

Indicator sequences

DNA AACTGGCATCCGGGAATAAGGTC

$x_A(n)$ 1 1 0 00 0 0 10 0 00 0 0 1 10 1 1 0 0 00

Indicator sequence for base A

Similarly define $x_T(n)$ $x_C(n)$ $x_G(n)$

$$\mathbf{x_A(n) + x_T(n) + x_C(n) + x_G(n) = 1}$$

Fourier transforms:

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

Fourier transforms:

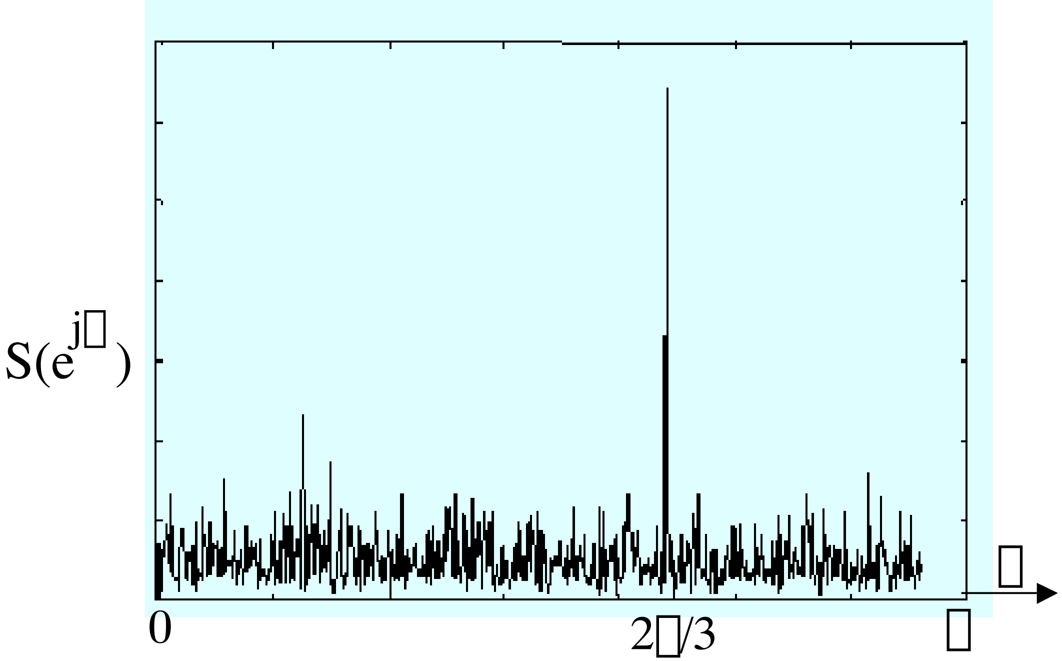
$$X_A(e^{j\omega}) \quad X_T(e^{j\omega}) \quad X_C(e^{j\omega}) \quad X_G(e^{j\omega})$$

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!

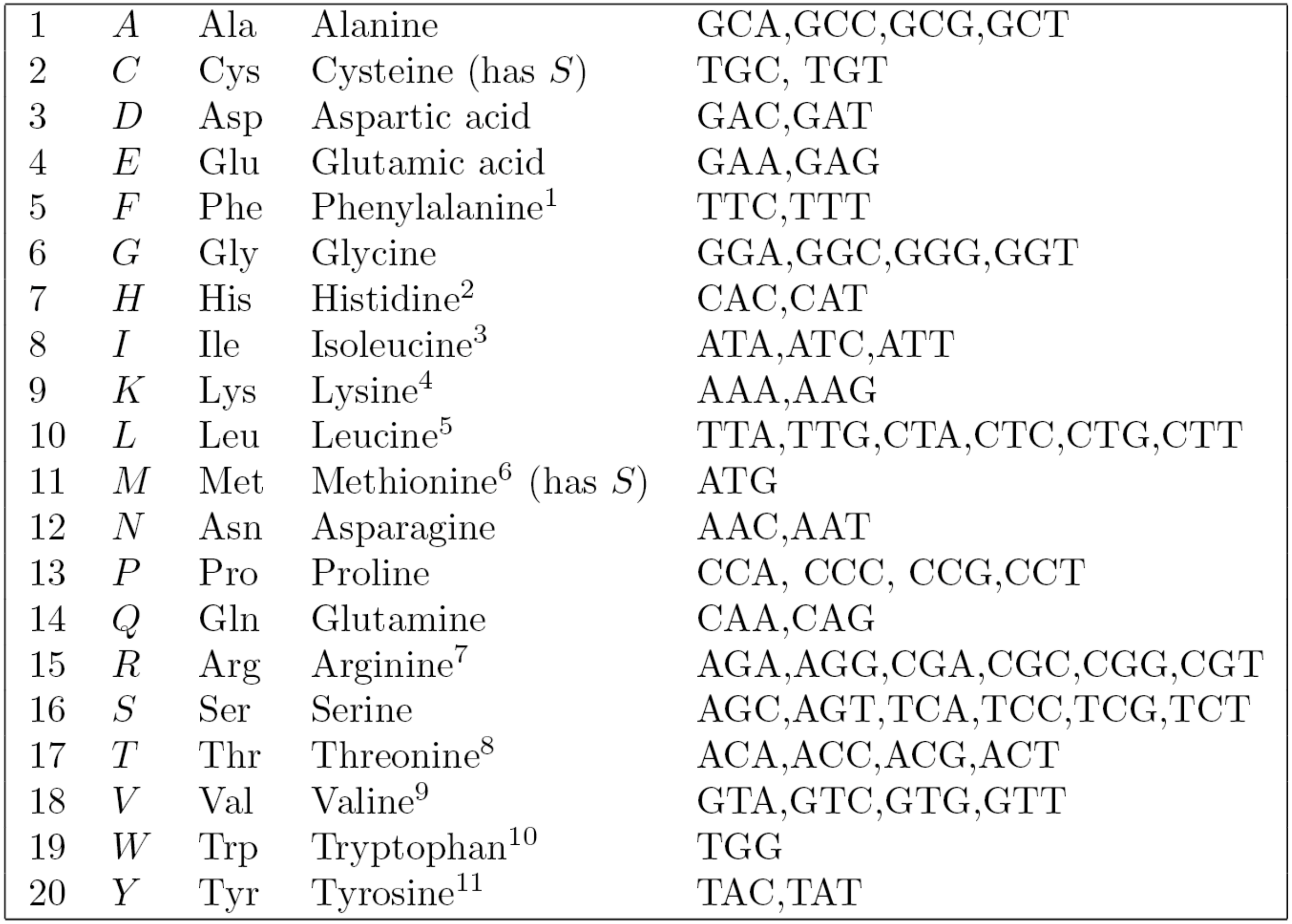


Coding region of length N=1320 inside a genome of baker's yeast (S. cerevisiae).

Tiwari, et. al., CABIOS, 1997.

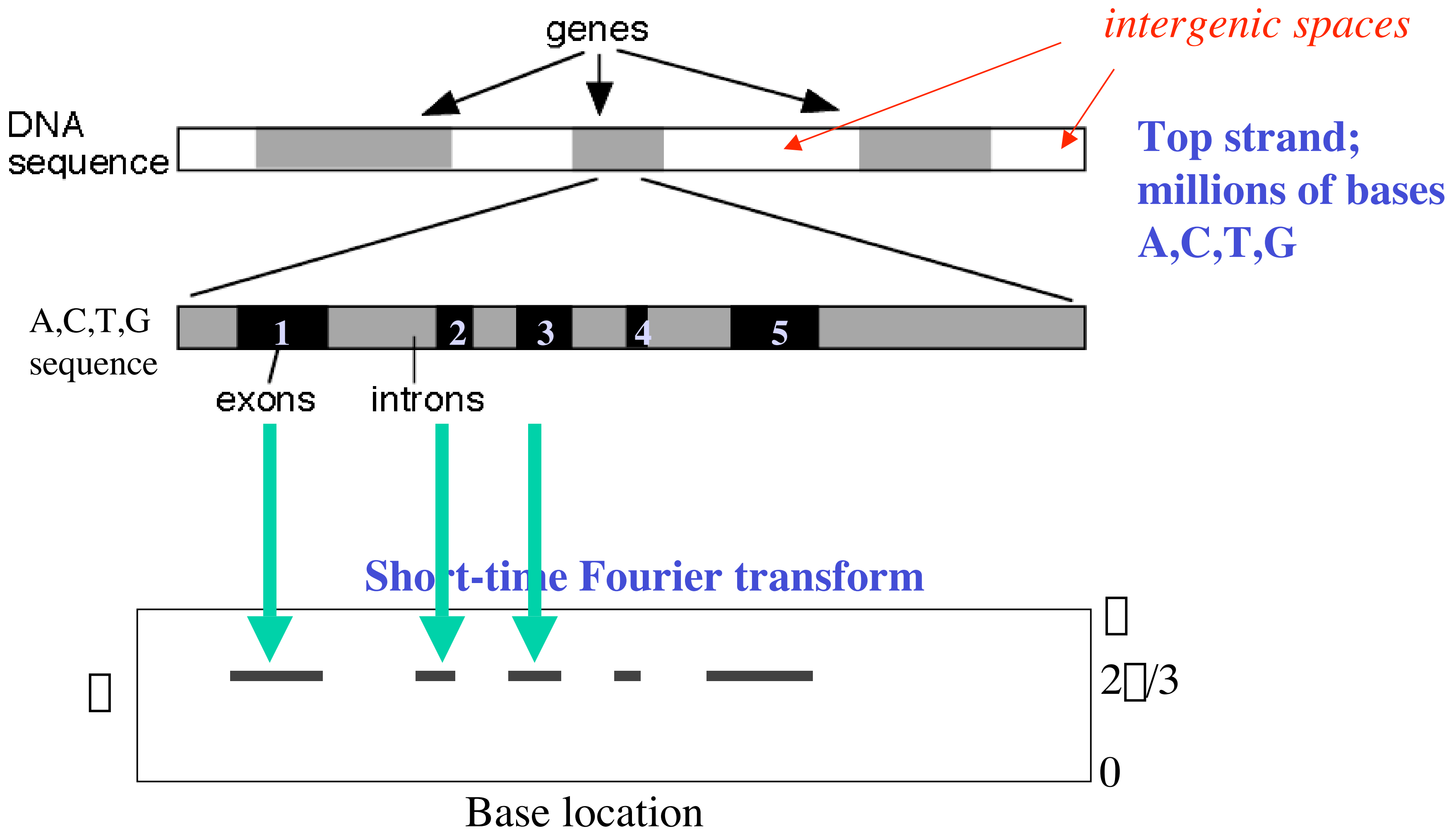
Dimitris Anastassiou, IEEE Signal Processing Magazine, July 2001

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.



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

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



intergenic spaces

Top strand; millions of bases A,C,T,G

Short-time Fourier transform

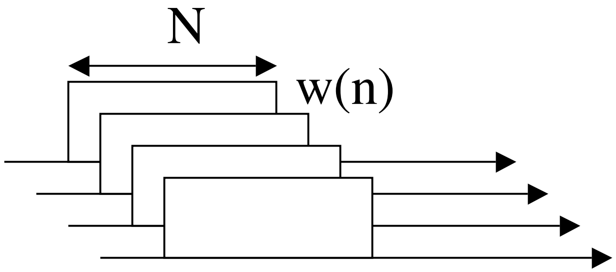
So we can locate exons using STFT

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

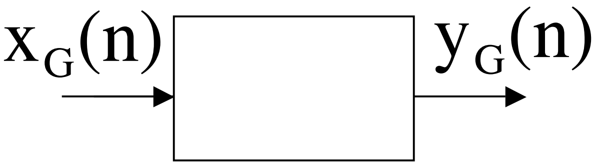
Filtering interpretation

Take any base, say G:

$x_G(n)$ 1 1 0 00 0 0 10 0 00 0 0 1 10 1 1 0 0 00 0 1 10 1 1 0 0 1 10 1 1 0

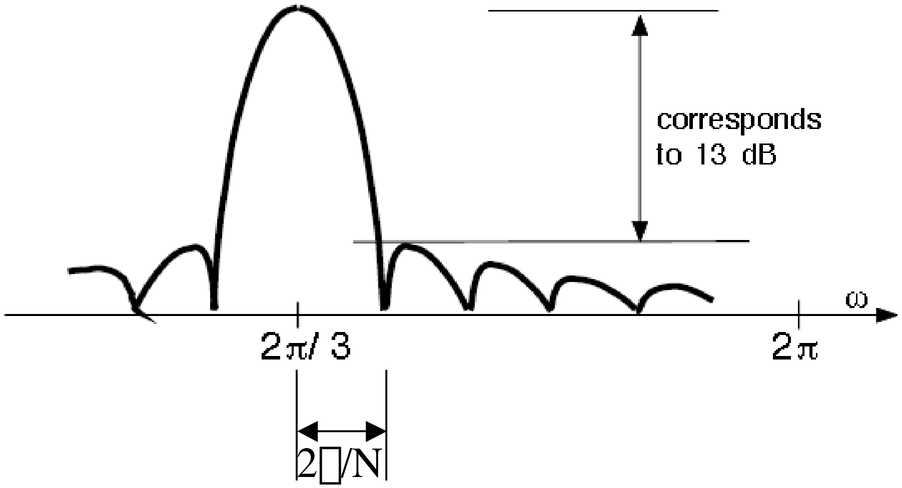


Sliding window

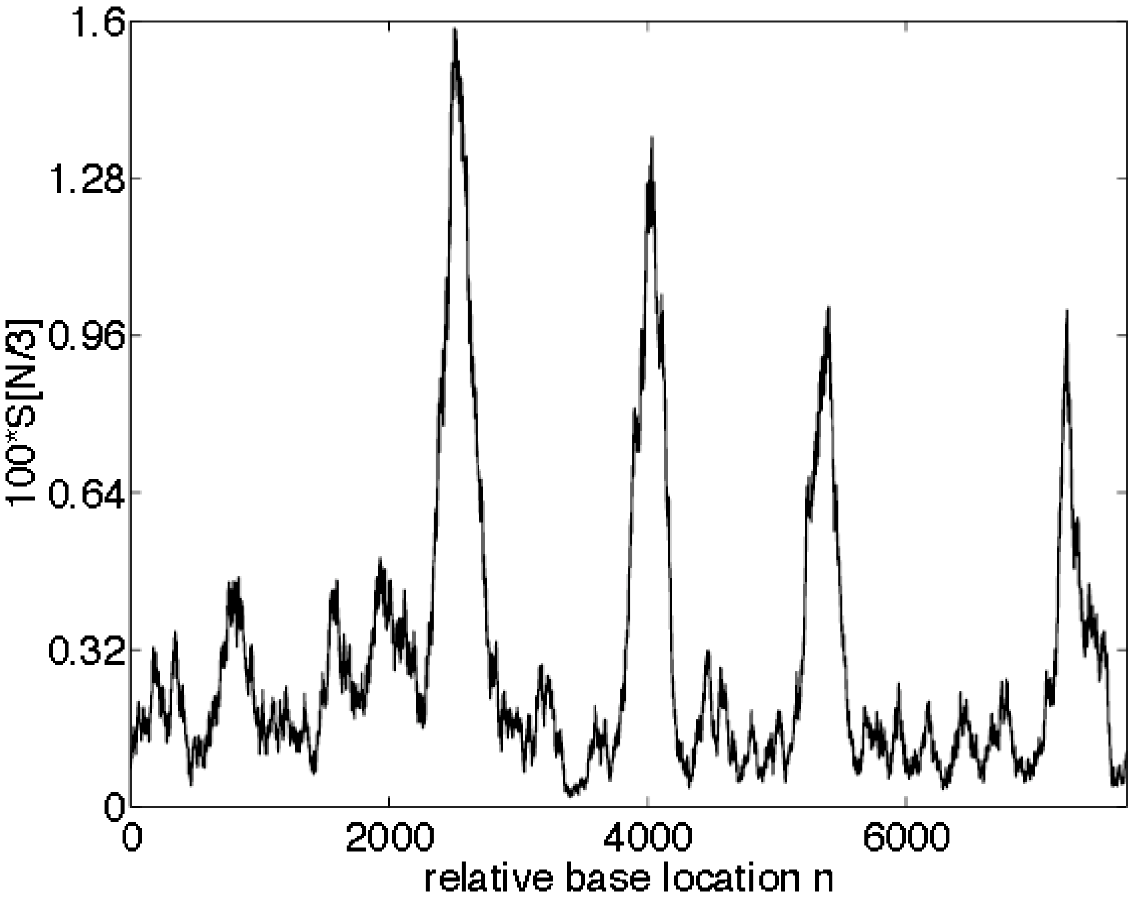


filter with impulse response h(n)

Frequency response magnitude



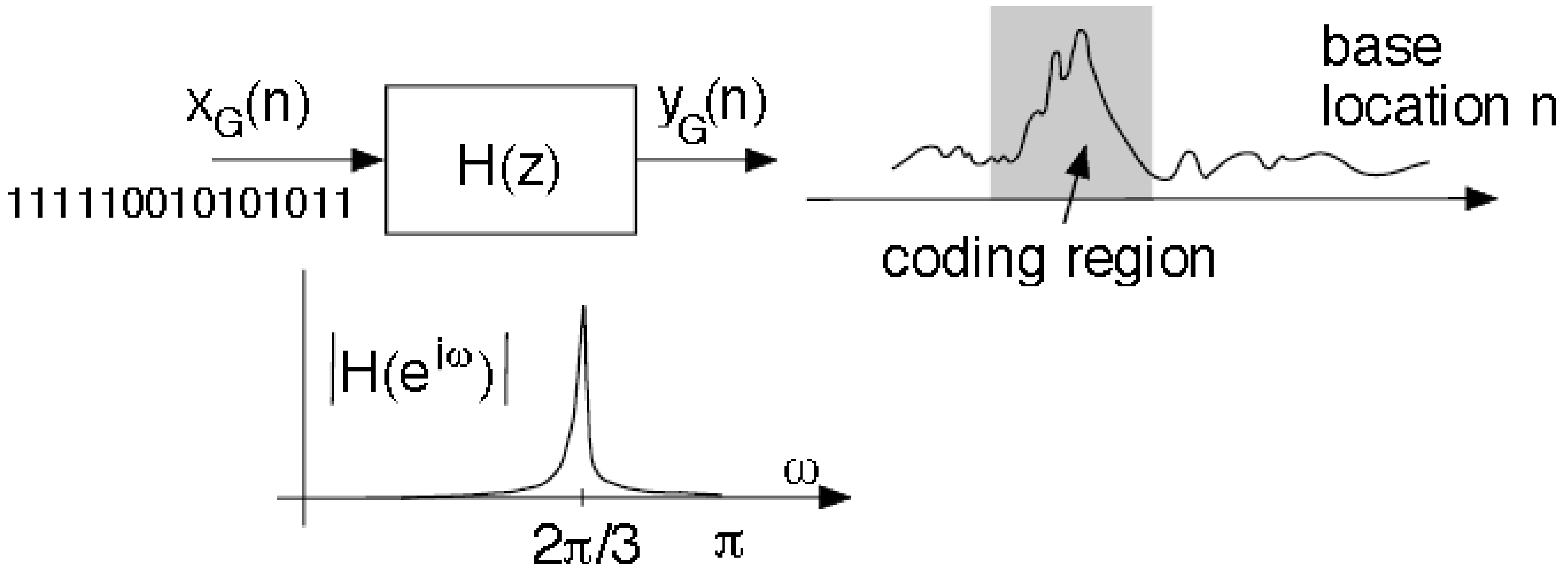
Spectrum at $2\pi/3$ as a function of base location



Gene F56F11.4 in the C-elegans chromosome III

Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

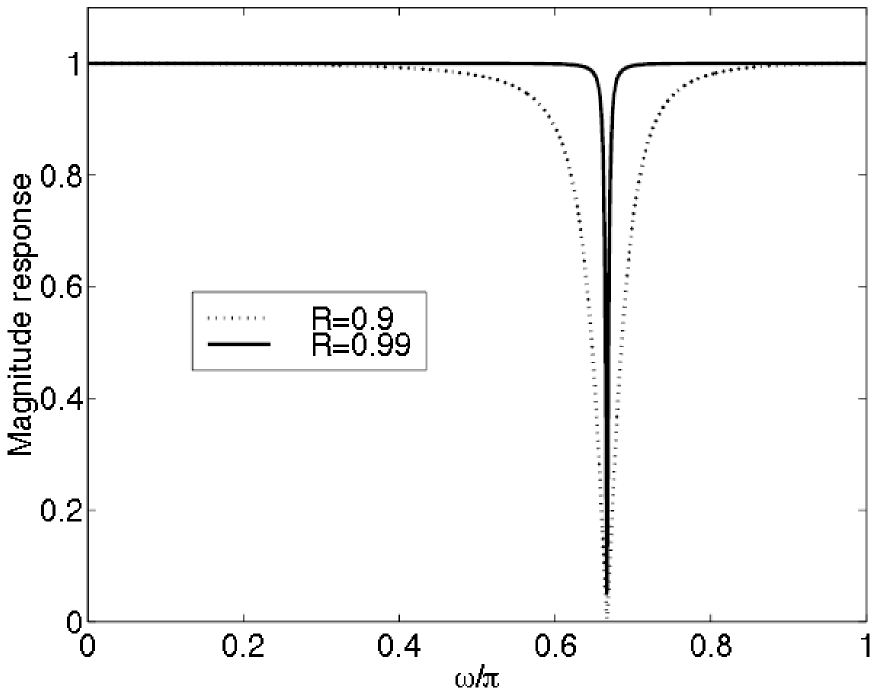
Return to the filtering interpretation



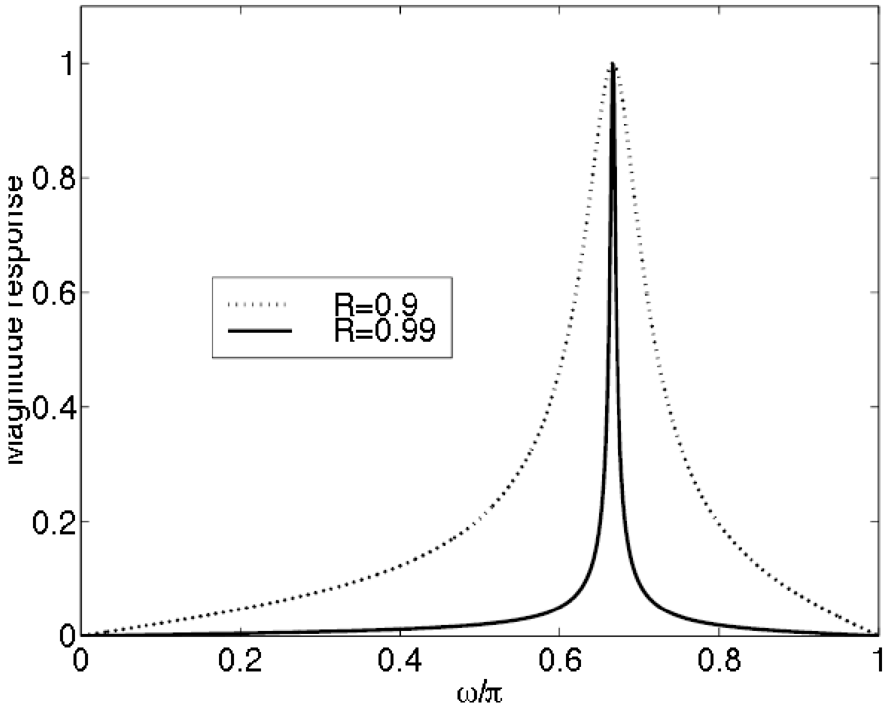
How about designing filters to improve time-frequency resolution?

Interesting DSP problem!

Notch



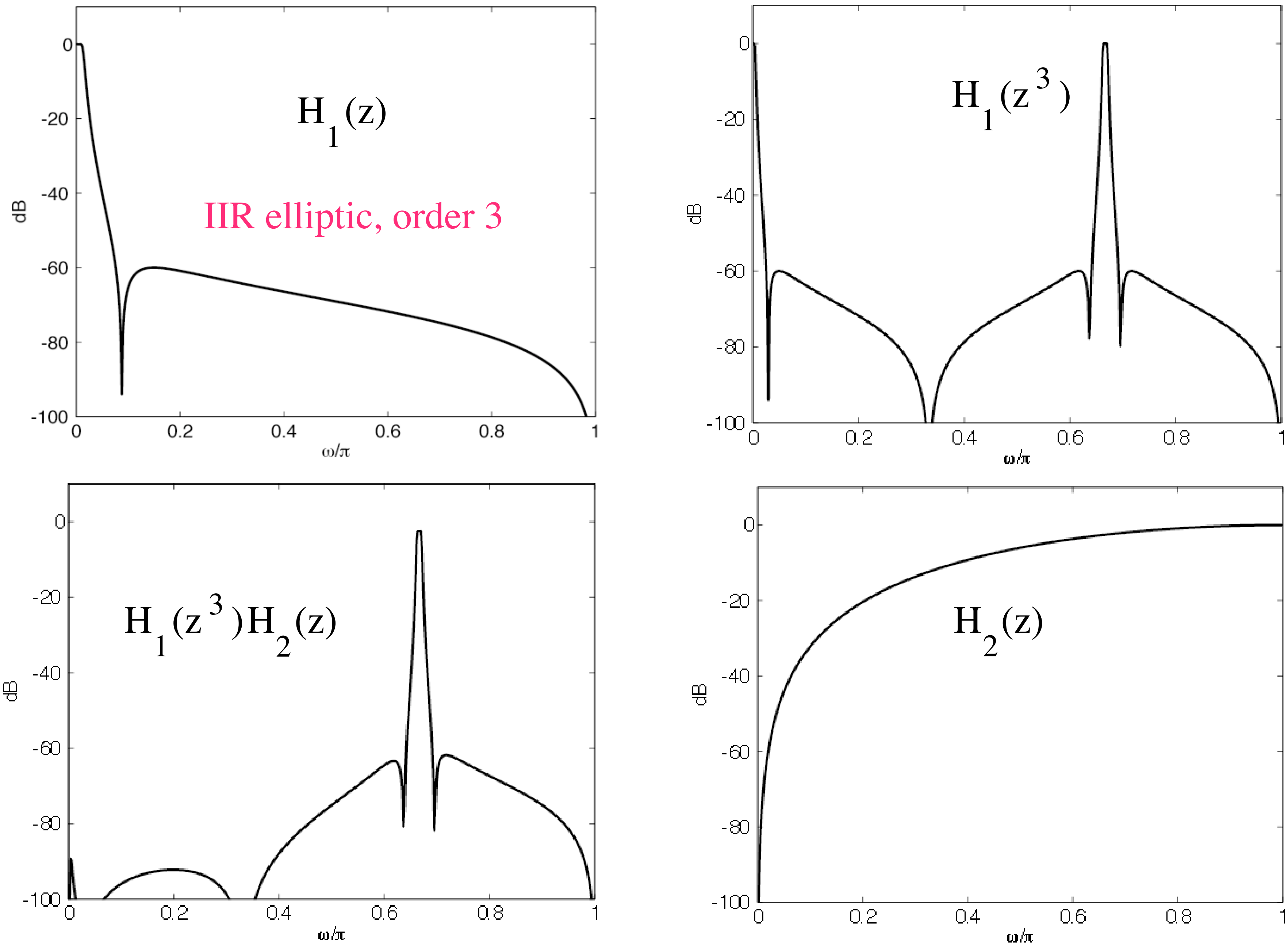
Antinotch



Allpass: $A(z) = \dfrac{R^2 - 2R\cos\theta\, z^{-1} + z^{-2}}{1 - 2R\cos\theta\, z^{-1} + R^2 z^{-2}}$

Define two filters: $\begin{bmatrix} G(z) \\ H(z) \end{bmatrix} = \dfrac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ A(z) \end{bmatrix}$

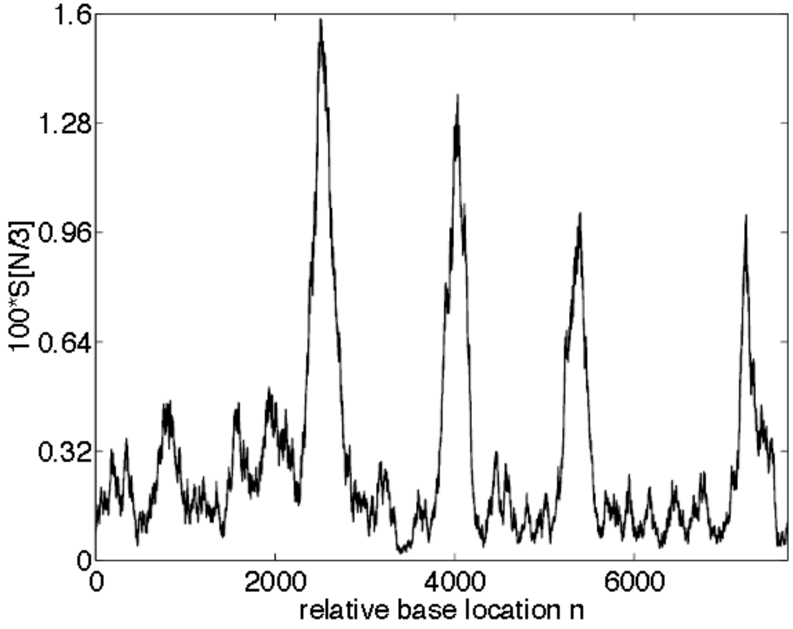
Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.



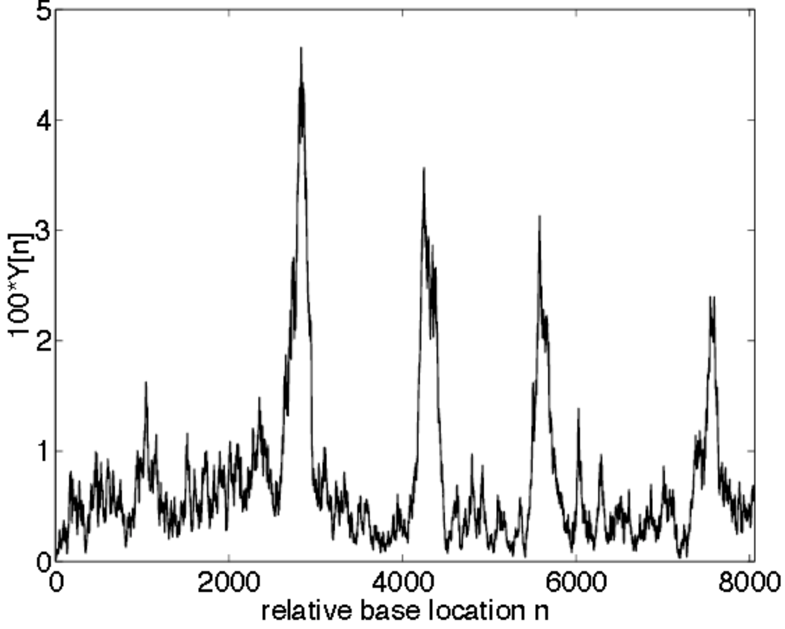
Multistage filter design method
like the IFIR method (Neuvo, et. al, 1983)

Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

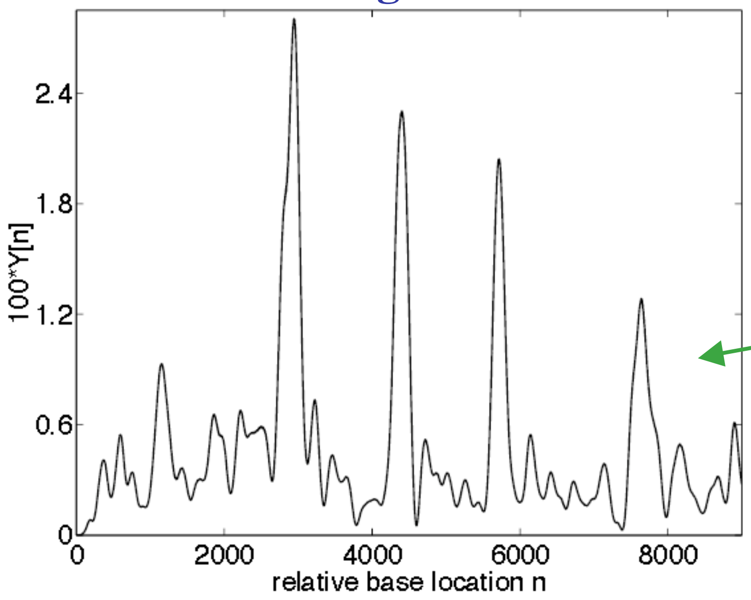
STFT calculation



Allpass based antinotch



Multistage antinotch



Sharper peaks

Low frequency noise removed

Gene F56F11.4 in the C-elegans chromosome III

Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

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

Will return to it later.

Outline

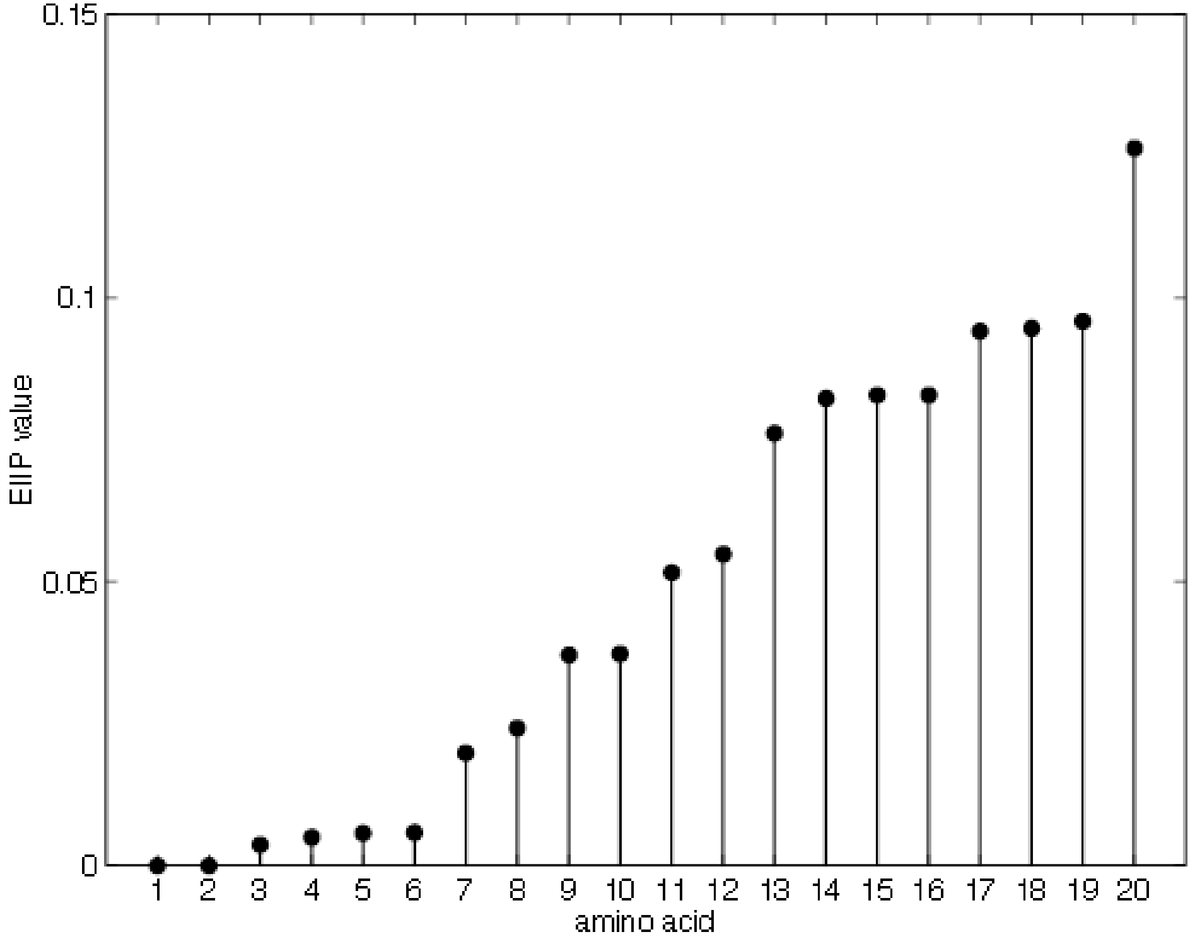
- ☺ Molecular biology background
- ☺ Computational gene-finding
- ☺ **Spectral analysis (Fourier, wavelet, correlations)**
- Hidden Markov Models and sequence analysis
- New world of non-coding genes
- References

Proteins are sequences made of 20 kinds of amino acids:

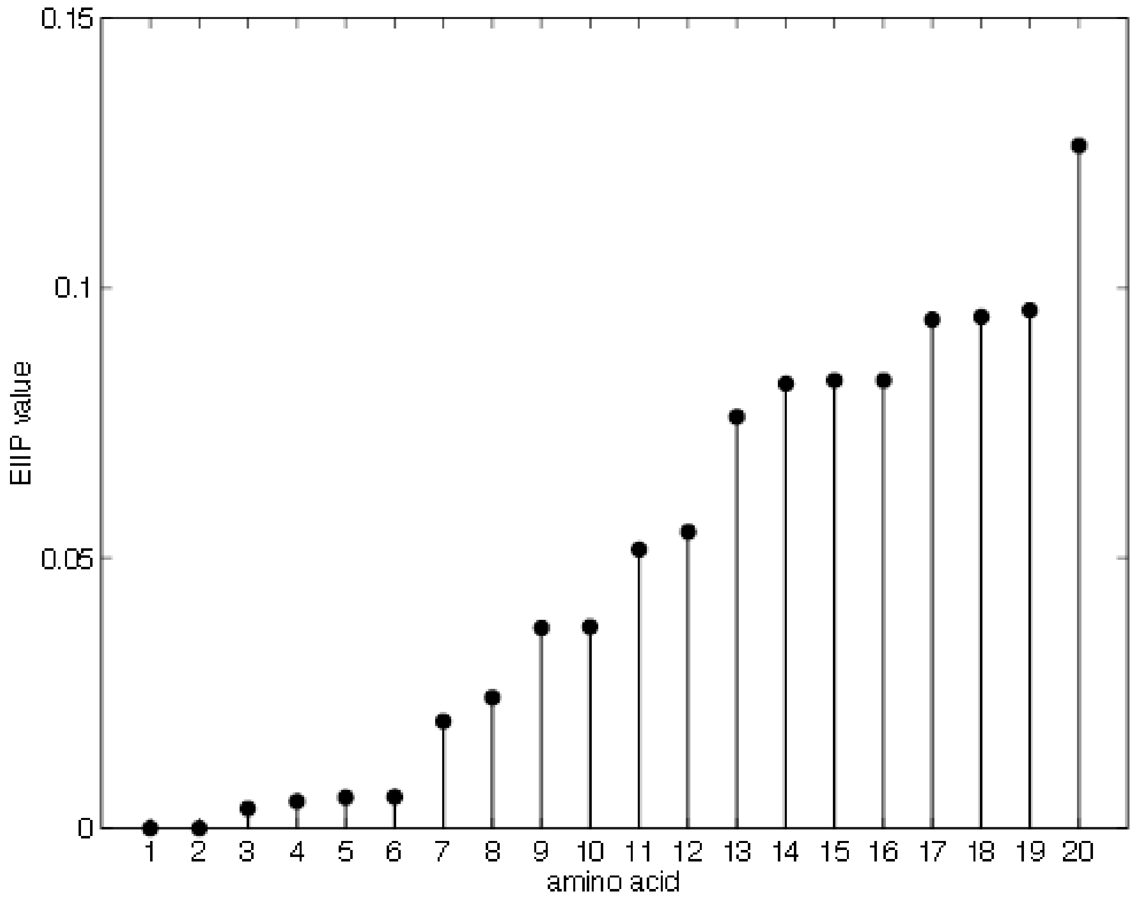
ACDEFGHIKLMNPQRSTVWY

Each amino acid is associated with a unique number called the **EIIP:**

Electron-ion interaction potential



I. Cosic, IEEE Trans. Biomed. Engr., Dec. 1994



Given an amino acid sequence: AACDEQRIKLYXTSVDC ....

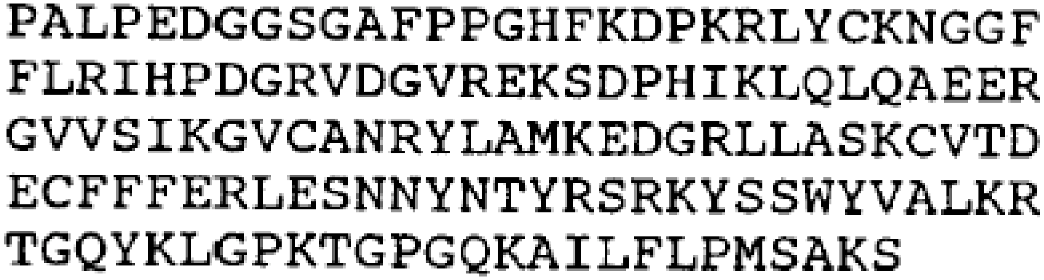
We can readily turn it into a numerical sequence x(n).

The Fourier transform of x(n) has interesting properties

Proteins belonging to the same functional group have something common in their Fourier transform!

Example: Fibroblast growth factor proteins

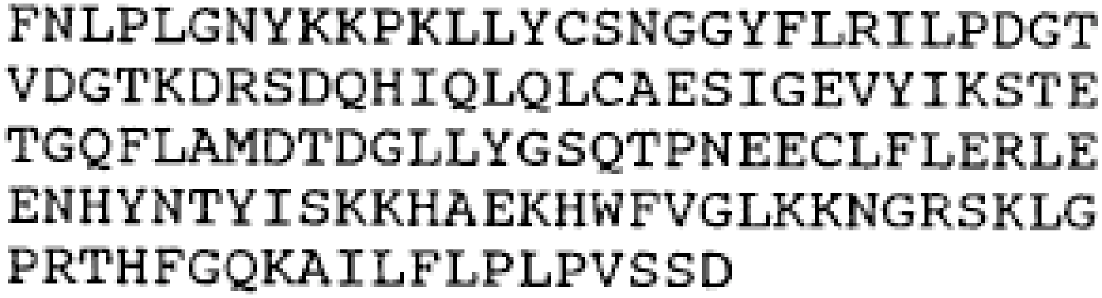
Basic bovine



```
PALPEDGGSGAFPPGHFKDPKRLYCKNGGF
FLRIHPDGRVDGVREKSDPHIKLQLQAEER
GVVSIKGVCANRYLAMKEDGRLLASKCVTD
ECFFFERLESNNYNTYRSRKYSSWYVALKR
TGQYKLGPKTGPGQKAILFLPMSAKS
```

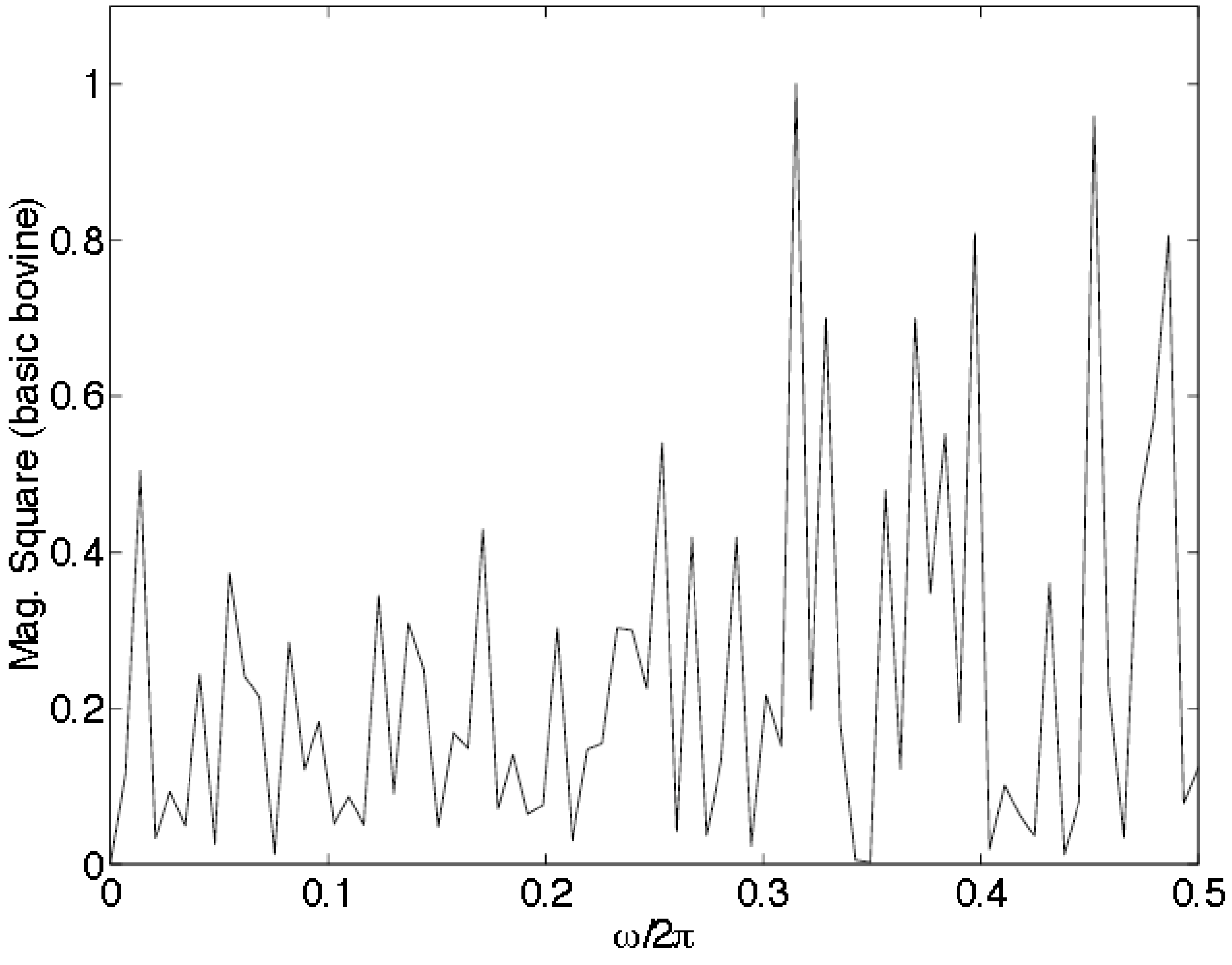
length 146

Acidic bovine

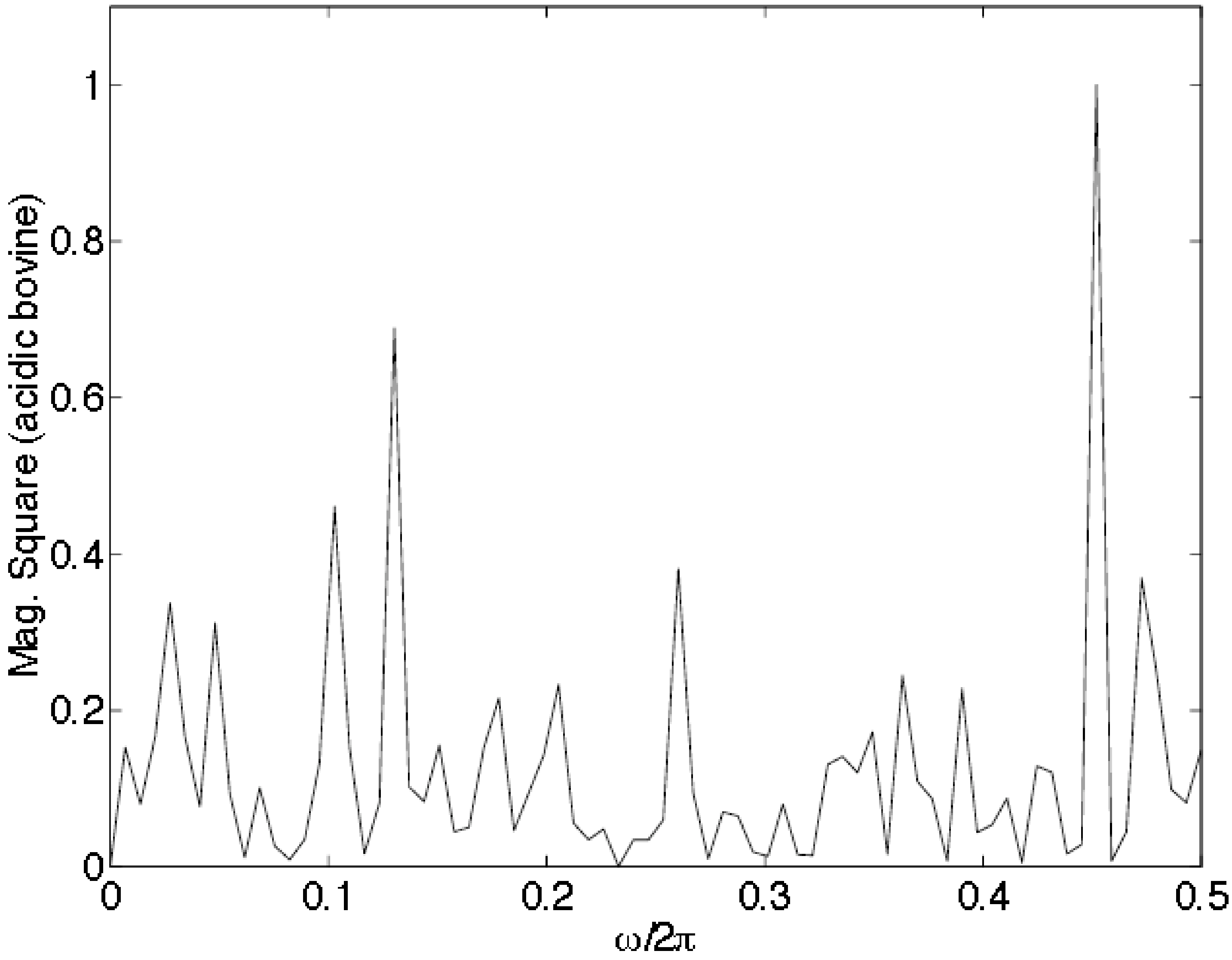


```
FNLPLGNYKKPKLLYCSNGGYFLRILPDGT
VDGTKDRSDQHIQLQLCAESIGEVYIKSTE
TGQFLAMDTDGLLYGSQTPNEECLFLERLE
ENHYNTYISKKHAEKHWFVGLKKNGRSKLG
PRTHFGQKAILFLPLPVSSD
```

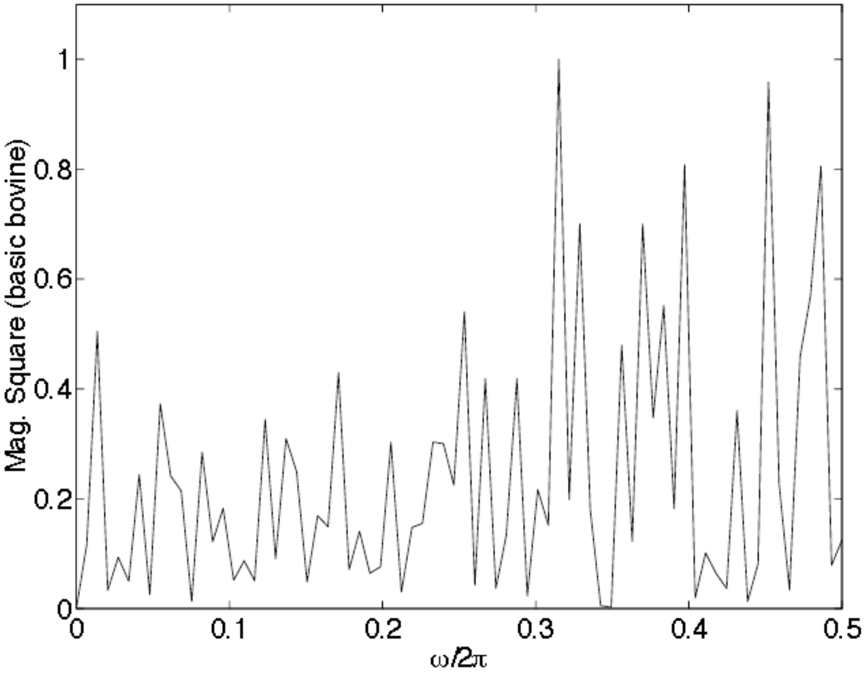
length 140

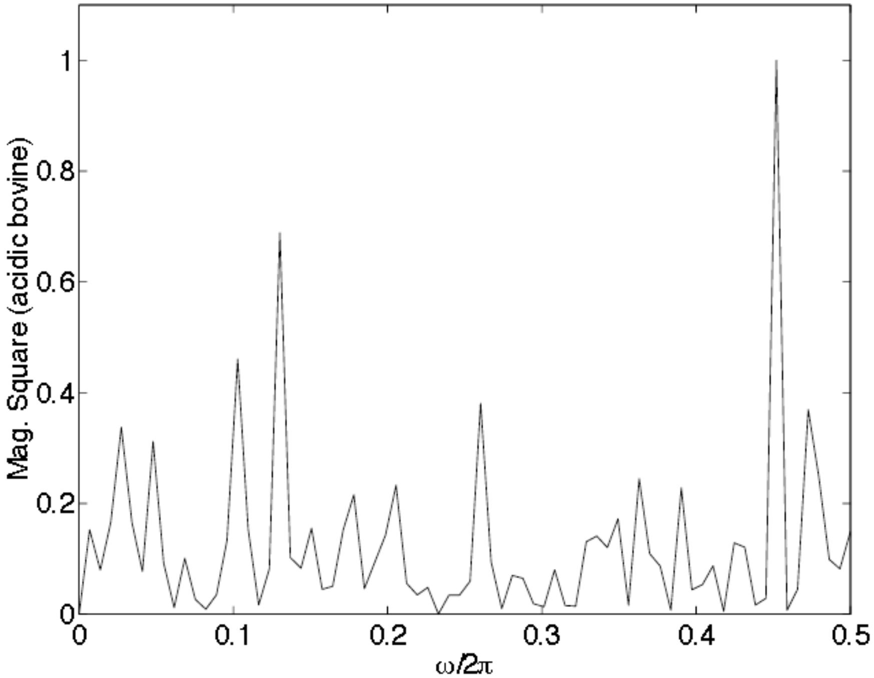


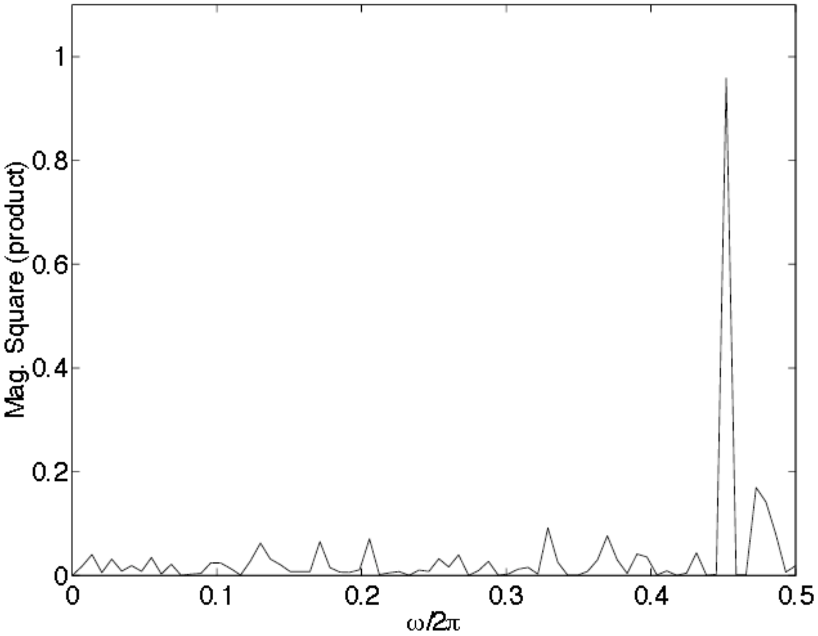
Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.



Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

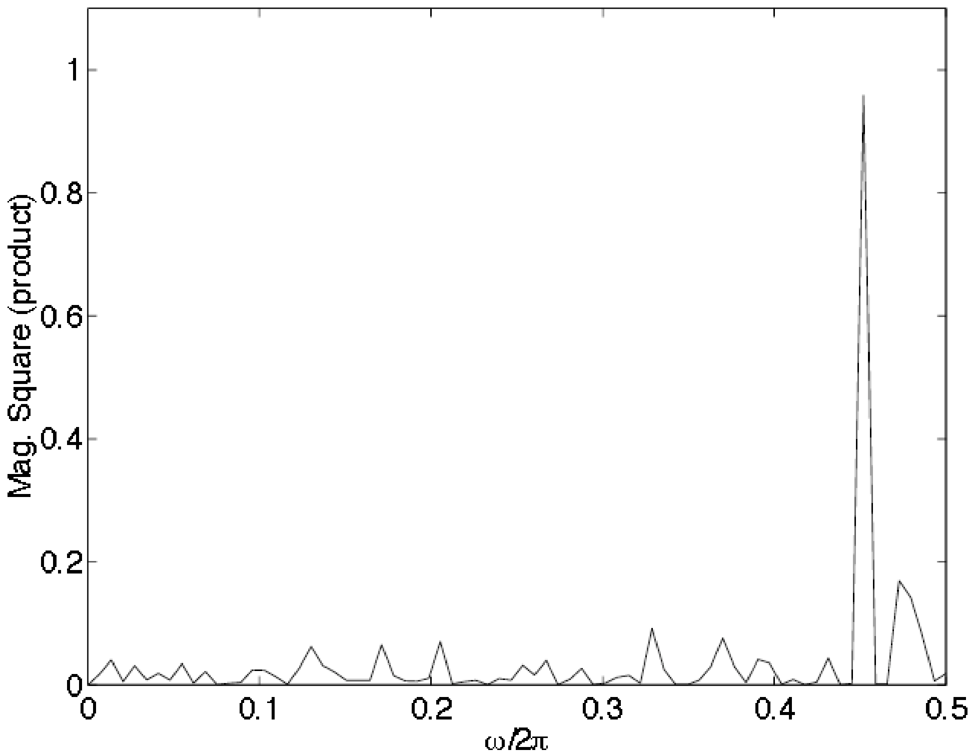






Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

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!

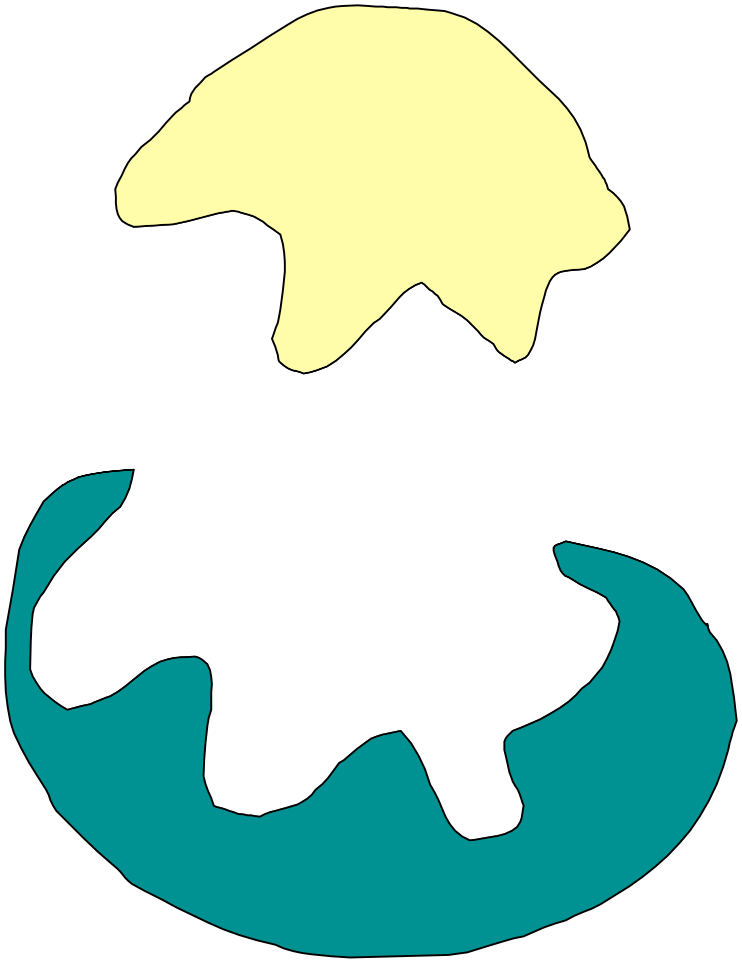


Proteins work by recognizing other molecules from **spatial periodic components!**

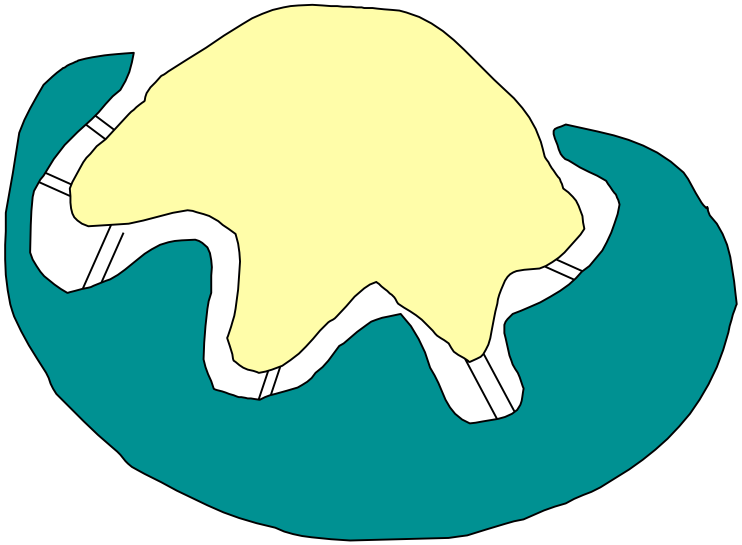
Resonant recognition model (RRM), Cosic, 1994.

Lots of good physics behind this. See references in Cosic, 1994.

some other molecule,
e.g., ligand

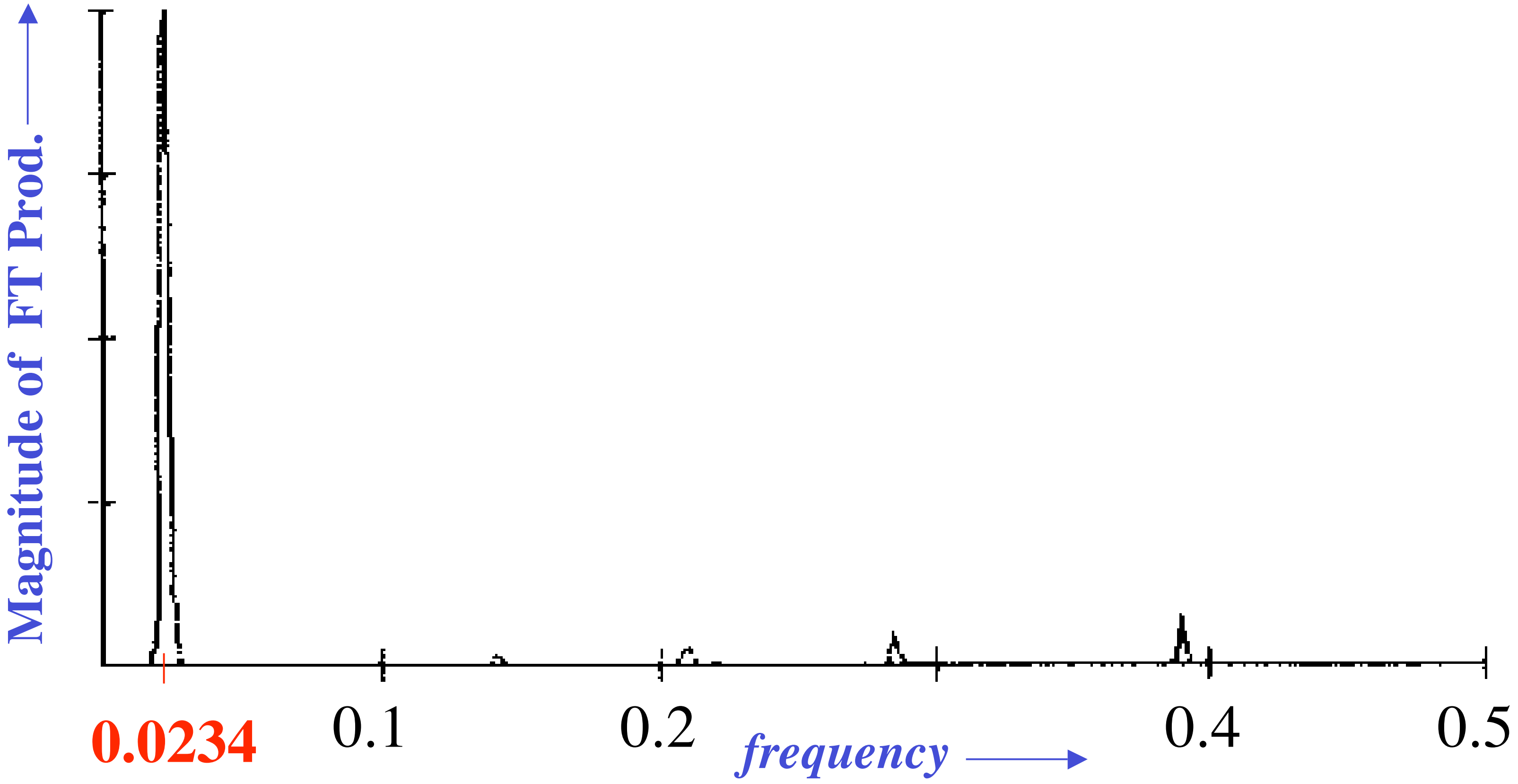


protein molecule



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

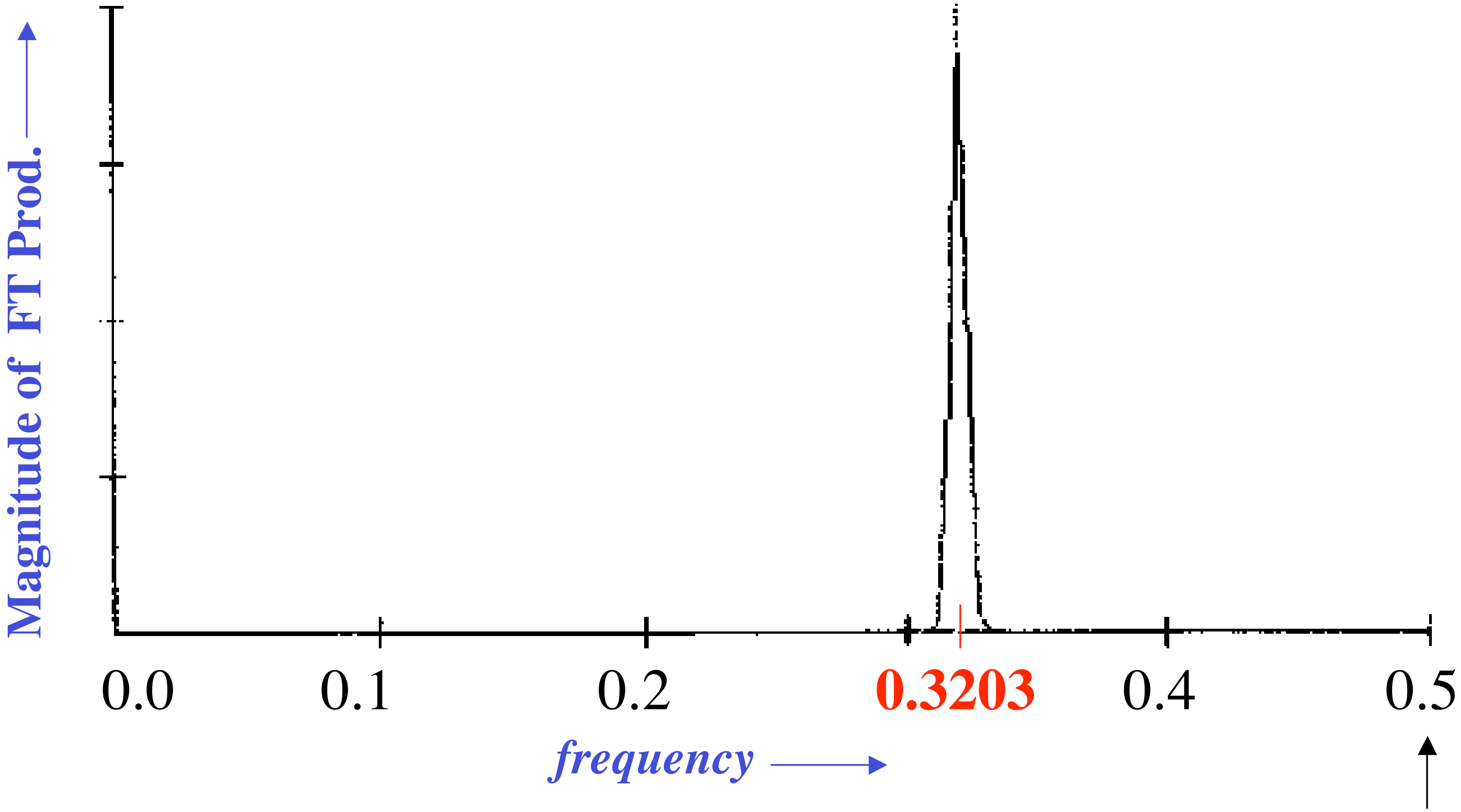
Protein group: hemoglobins



Adapted from Cosic, IEEE Trans. Biomed Engr., 1994.

Hemoglobins are oxygen carriers in the red blood cells.

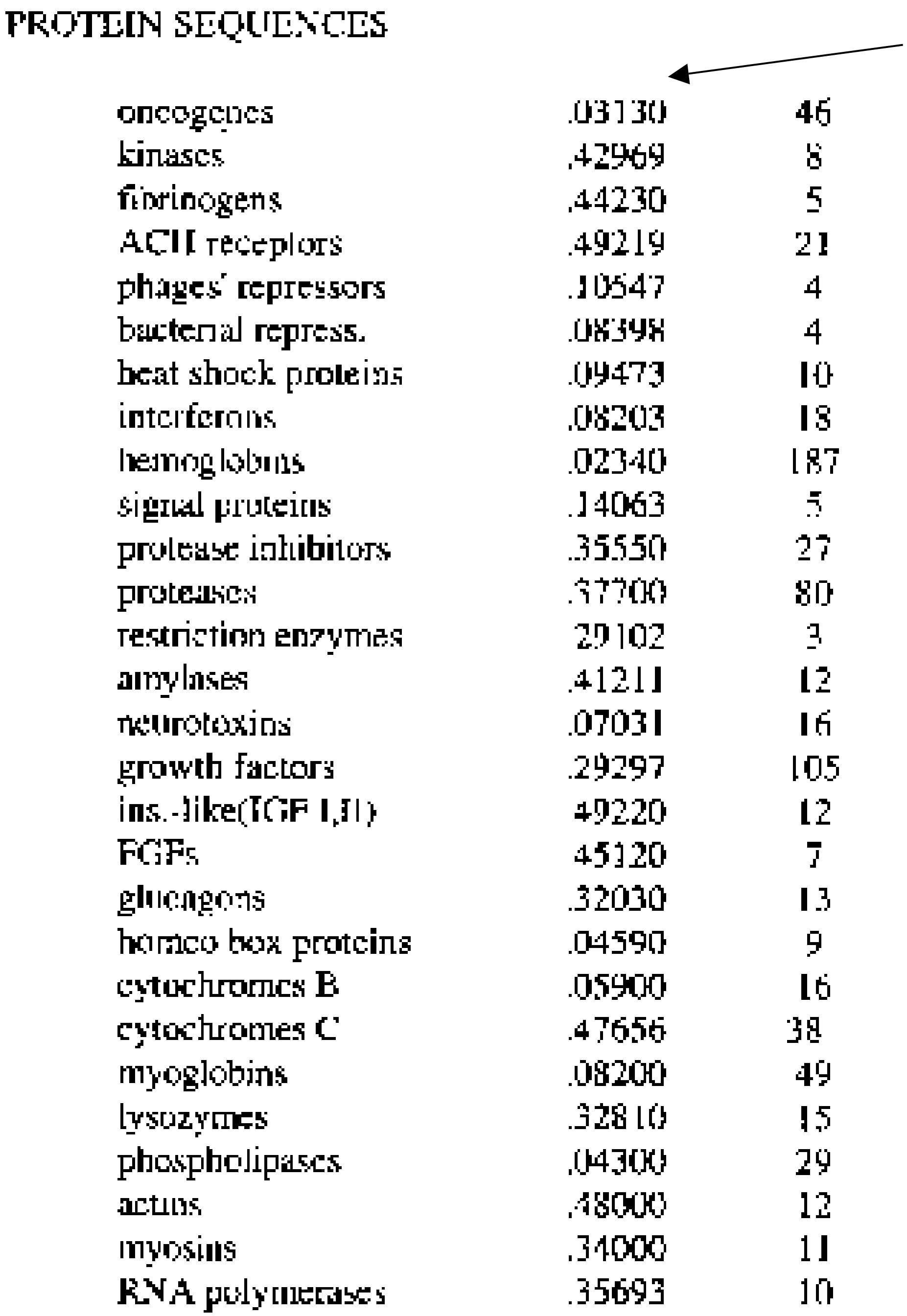
Protein group: glucagons



this is like □

Adapted from Cosic, IEEE Trans. Biomed Engr., 1994.

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



PROTEIN SEQUENCES

oncogenes	.03130	46
kinases	.42969	8
fibrinogens	.44230	5
ACH receptors	.49219	21
phages' repressors	.10547	4
bacterial repress.	.08398	4
heat shock proteins	.09473	10
interferons	.08203	18
hemoglobins	.02340	187
signal proteins	.14063	5
protease inhibitors	.35550	27
proteases	.37700	80
restriction enzymes	29102	3
amylases	.41211	12
neurotoxins	.07031	16
growth factors	.29297	105
ins.-like(IGF I,II)	49220	12
FGFs	45120	7
glucagons	.32030	13
homeo box proteins	.04590	9
cytochromes B	.05900	16
cytochromes C	.47656	38
myoglobins	.08200	49
lysozymes	.32810	15
phospholipases	.04300	29
actins	.48000	12
myosins	.34000	11
RNA polymerases	.35693	10

Frequency normalized so that 2π corresponds to 1

Examples of other functional groups of proteins.

Cosic, IEEE Trans. Biomed Engr., 1994.

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

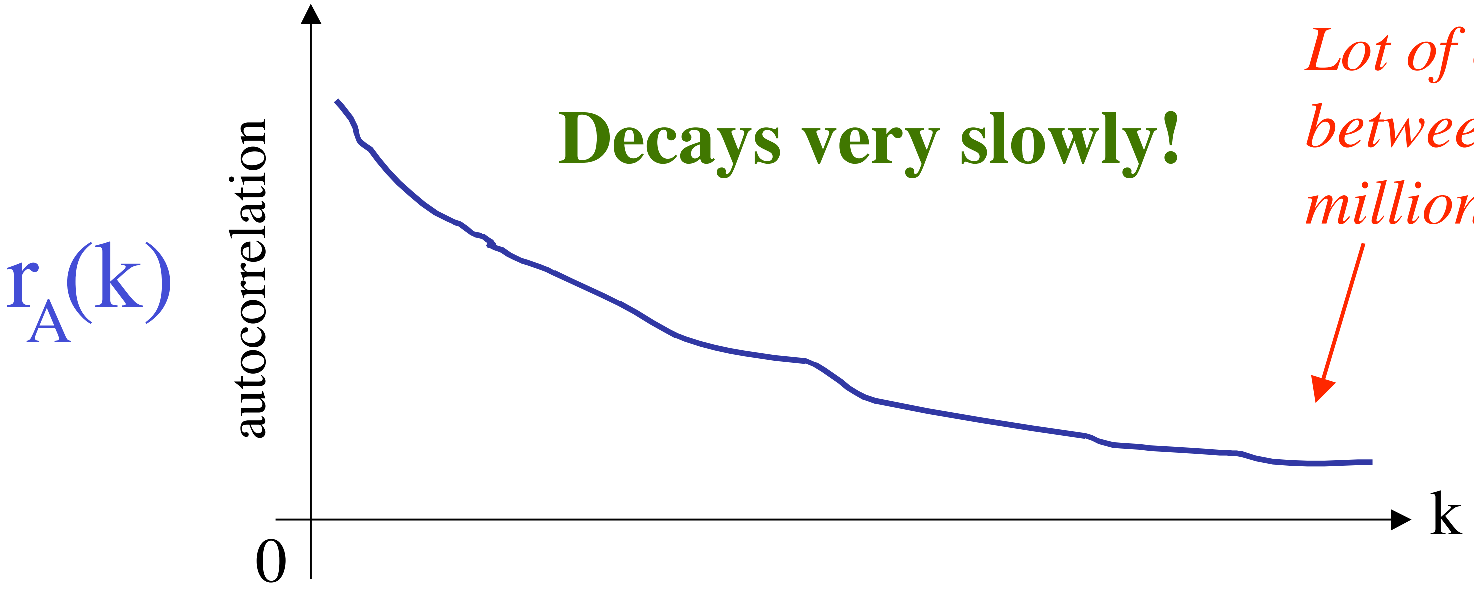
- Usual tradeoff between frequency localization and time localization.
- **Wavelet transform**: natural candidate for this.

Piragova, et al., Proc. of the IEEE, Dec. 2002.

Long-range correlation in DNA sequences

DNA AACTGGCATCCGGGAATAAGGTC

$x_A(n)$ 1 1 0 00 0 0 10 0 00 0 0 1 10 1 1 0 0 00



Decays very slowly!

Lot of correlation between bases millions away

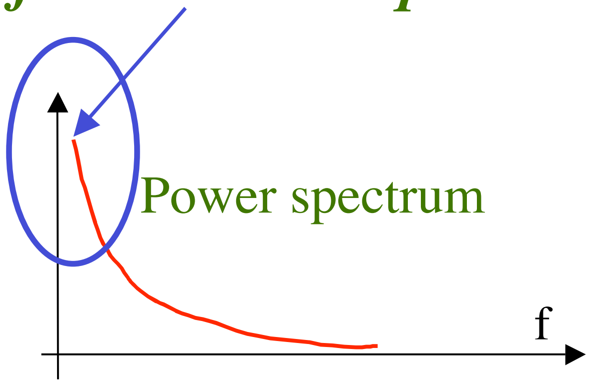
Long-range correlation or 1/f property

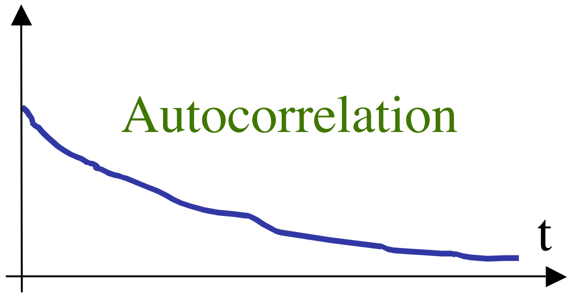
Fourier transform pair:

$$\frac{1}{|f|^{\alpha}} \Leftrightarrow c\,|t|^{\alpha-1}$$

called **1/f property** for any □ > 0.

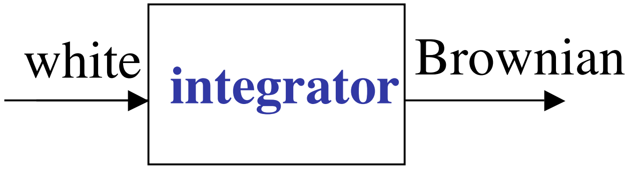
1/f behavior is equivalent to long range correlation in time.





Examples:

- □ =1 for traditional 1/f noise.
- □ = 2 for Brownian noise.



white → **integrator** → Brownian

Papoulis, Systems and transforms, 1968

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

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).

de Sousa Vieira, Physical review E, Nov. 99 (studied many organisms).

Li, Physical review A, May 1991 (duplicate-mutate theory).

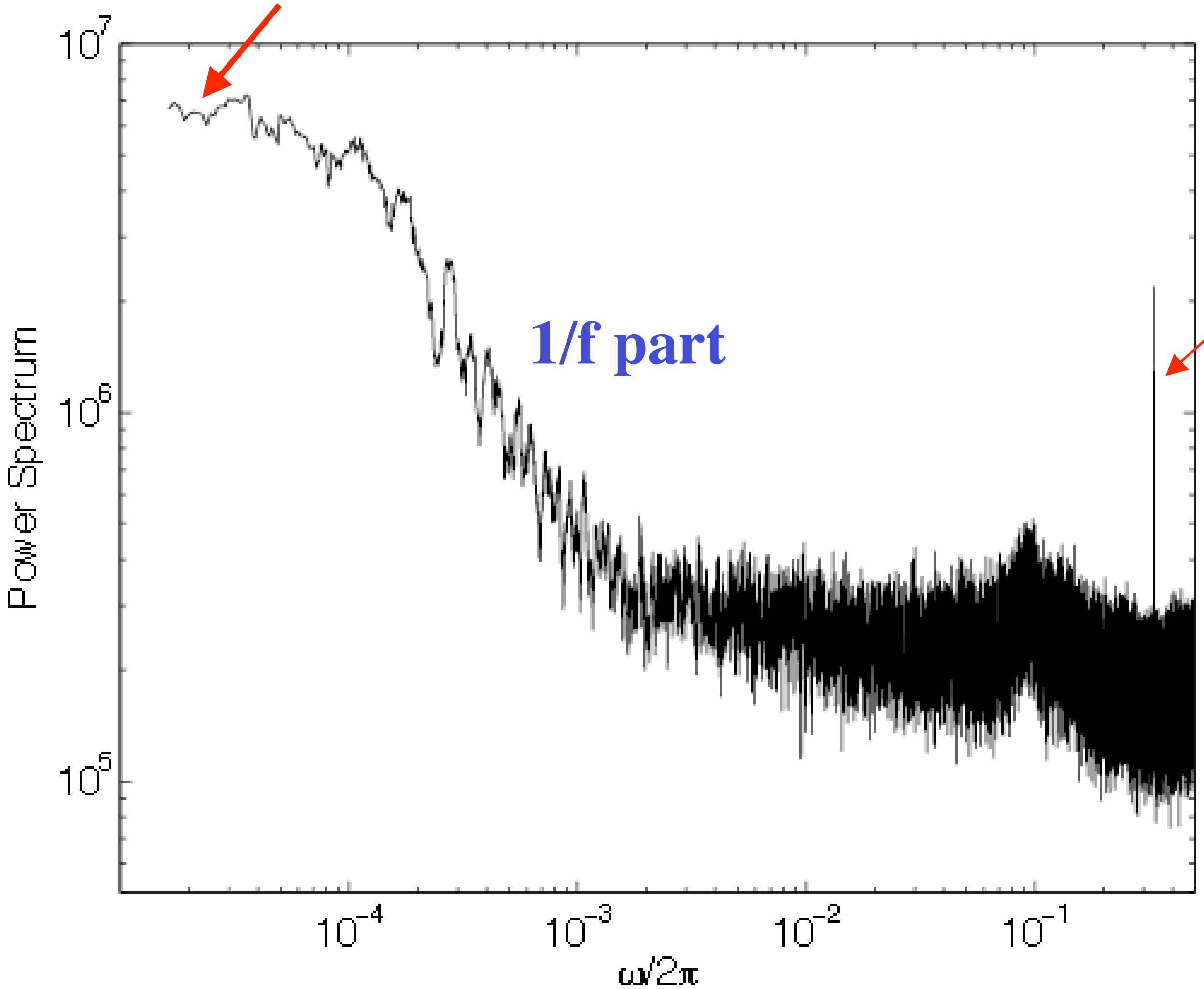
Hausdroff and **Peng**, Physical review E, Aug. 96 (multiscale randomness).

Early work on theory:

Wornell, IEEE Trans. IT, July 1990: 1/f noise modeled using with wavelets.

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

Example: Bacteria aquifex aeolicus, size 1.55 Mb.



flattens out

1/f part

line at
$\omega = 2\pi/3$
(period-3 component)

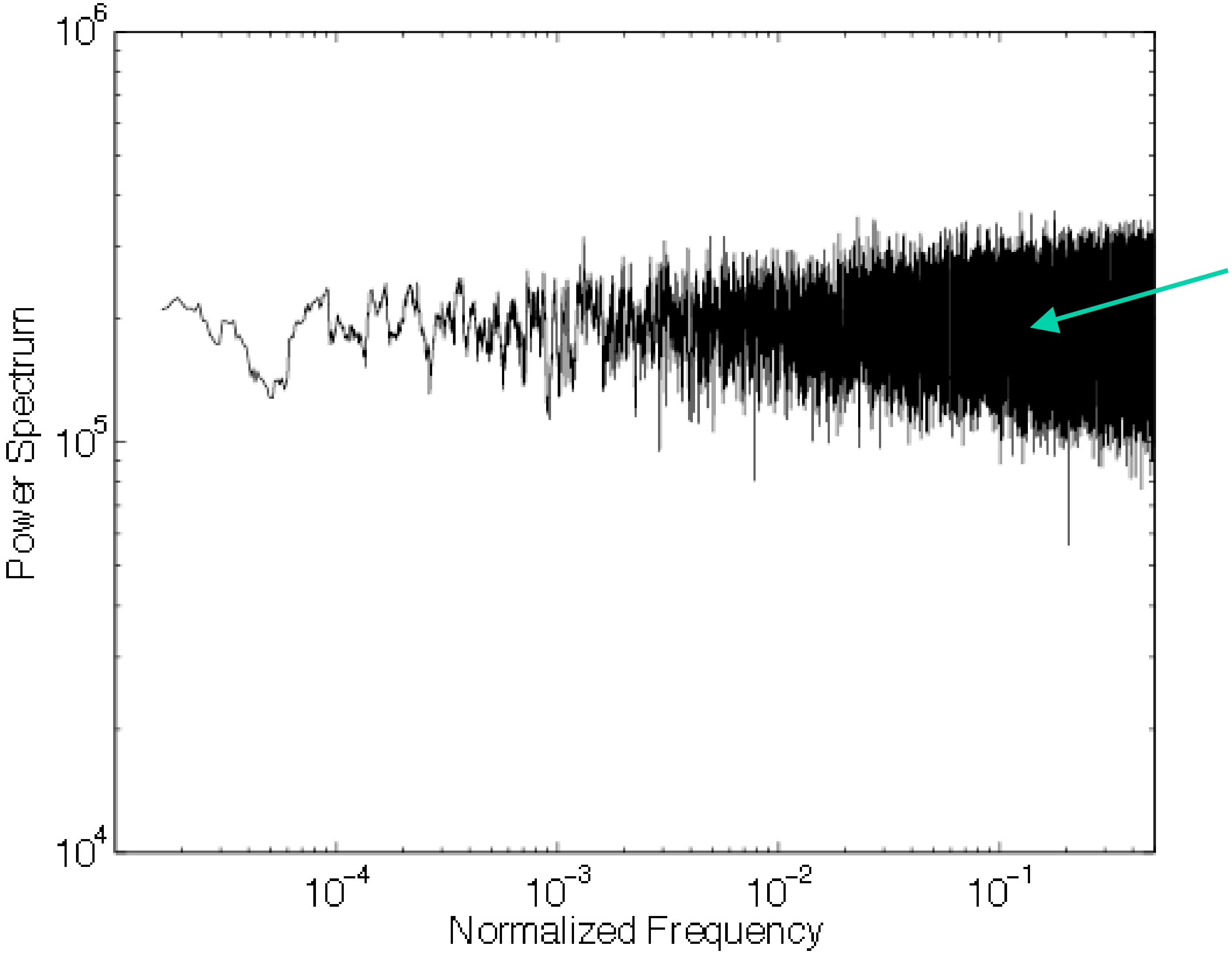
This is a typical DNA power spectrum

PSD for base A; 1 million bases used

de Sousa Vieira, 1999

Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

PSD of base A in randomly generated “DNA”.



Thickening due to log-log axis

No evidence of any 1/f behavior

Vaidyanathan and Yoon, J. of the Franklin Inst., Elsevier Ltd., 2004.

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

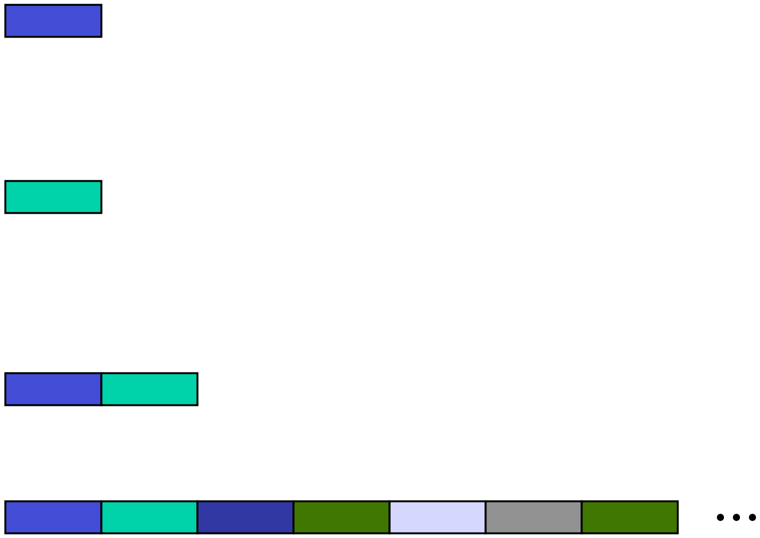
- Earliest life: few **1000** bases (**half a billion years ago**)
- Today's smallest bacteria : few **million** bases
- Mammals like us: few **billion** bases.

Evolution model: **duplicate and mutate** model.

Mathematical challenge

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

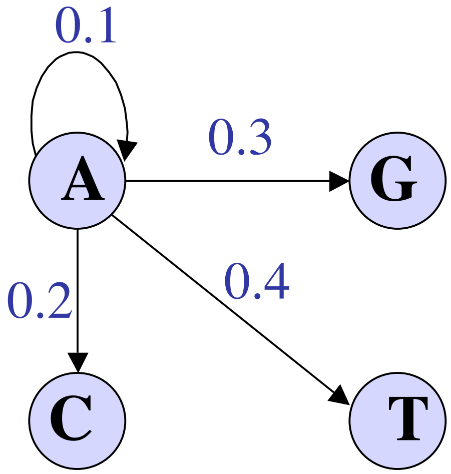
End of this part

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*:** **AA**CTG**AG**GT**AC**AATTCG**AT**CTC



$$\begin{array}{c|cccc} & \mathbf{A} & \mathbf{C} & \mathbf{T} & \mathbf{G} \\ \hline \mathbf{A} & 0.1 & 0.2 & 0.4 & 0.3 \\ \mathbf{C} & 0.2 & 0.5 & 0.1 & 0.2 \\ \mathbf{T} & 0.5 & 0.2 & 0.1 & 0.2 \\ \mathbf{G} & 0.3 & 0.1 & 0.4 & 0.2 \end{array}$$

State transition matrix □

L. R. Rabiner, Proc. IEEE, Feb. 1989.

Application of Markov models

Given a DNA sequence: **X =** x(1) x(2) x(3) …… x(N)
And given a Markov model □, we can calculate:

Probability that sequence X is generated by model □ :

P(**X**) = P(x(1)) x P(x(1) to x(2)) x P(x(2) to x(3)) x …..

Given a set of models: □$_1$ □$_2$ •••• □$_K$

Model 1 Model 2 Model K

exons introns intergenic

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

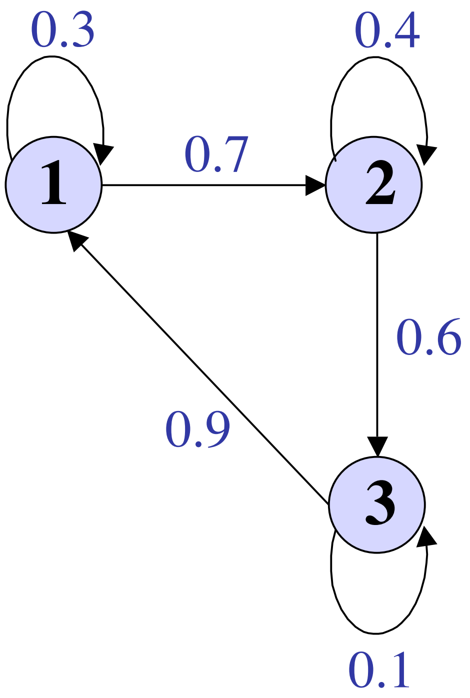
The models are obtained by training with known sequences.

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

Hidden Markov Models (HMM)

In an HMM, states are not the same thing as outputs.

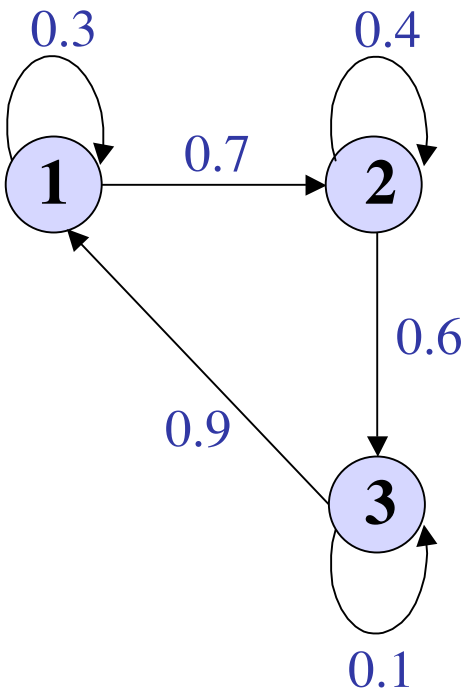
Example: States: 1, 2, 3 Outputs: A, C, T, G



State 1	State 2	State 3
A: 0.3	A: 0.5	A: 0.1
C: 0.1	C: 0.3	C: 0.3
T: 0.4	T: 0.1	T: 0.4
G: 0.2	G: 0.1	G: 0.2

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

HMM example:



A: 0.3
C: 0.1
T: 0.4
G: 0.2

State 1

A: 0.5
C: 0.3
T: 0.1
G: 0.1

State 2

A: 0.1
C: 0.3
T: 0.4
G: 0.2

State 3

$$\begin{array}{c|ccc|}
 & \mathbf{1} & \mathbf{2} & \mathbf{3} \\
\hline
\mathbf{1} & 0.3 & 0.7 & 0.0 \\
\mathbf{2} & 0.0 & 0.4 & 0.6 \\
\mathbf{3} & 0.9 & 0.0 & 0.1
\end{array}$$

State transition matrix □

$$\begin{array}{c|cccc|}
 & \mathbf{A} & \mathbf{C} & \mathbf{T} & \mathbf{G} \\
\hline
\mathbf{1} & 0.3 & 0.1 & 0.4 & 0.2 \\
\mathbf{2} & 0.5 & 0.3 & 0.1 & 0.1 \\
\mathbf{3} & 0.1 & 0.3 & 0.4 & 0.2
\end{array}$$

Output matrix □

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

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

Salzberg, Searls, and Kasif, *Computational methods in molecular biology, Elsevier, 1998.*

Durbin, Eddy, Krogh, and Mitchison, *Biological sequence analysis,Cambridge Univ. Press, 1998.*

HMM is a finite state machine (FSM) and represents **regular grammars.**

Regular grammar

Only production-rules of the form: W → aW

W: nonterminal *a: terminal*

Example: suppose the grammar is defined by these rules:

W → AW W → TW W → CW

Example of a string generated by this grammar:

W → AW → AAW → AACW → AACTW → AACT

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

Stochastic means: each rule is used with a certain probability

Regular grammar example:

W → AW → AAW → AACW → AACTW → AACT

Context free grammar (CFG):

Production rules of the form: W → □

W: nonterminal *□: string of terminal and or nonterminals*

Example: grammar with production rules:

W → AWA W → CWC W → TWT W → GWG W → null

Example of sequence generated:

W → AWA → ATWTA → ATCWCTA → ATCCTA

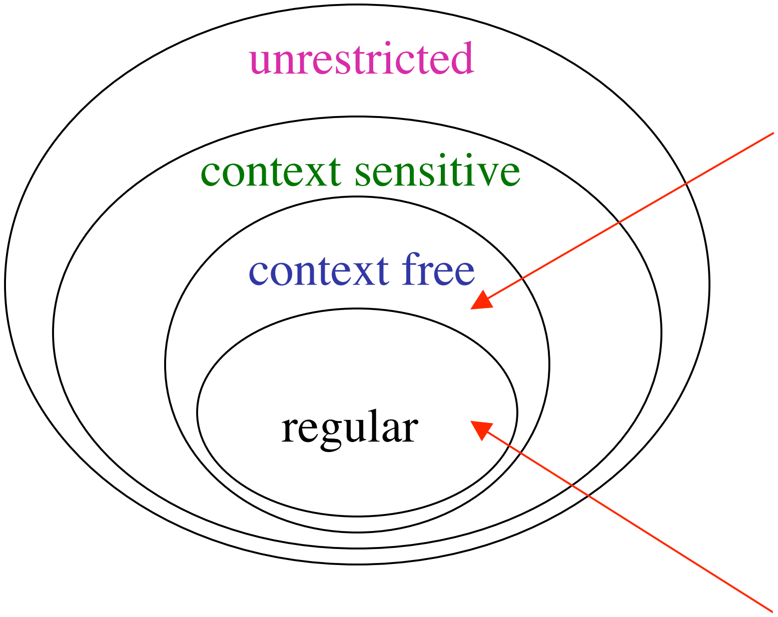
This is a symmetric sequence (palindrome)

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)



SCFG non-coding genes
ncRNAs
siRNAs

these have palindrome components

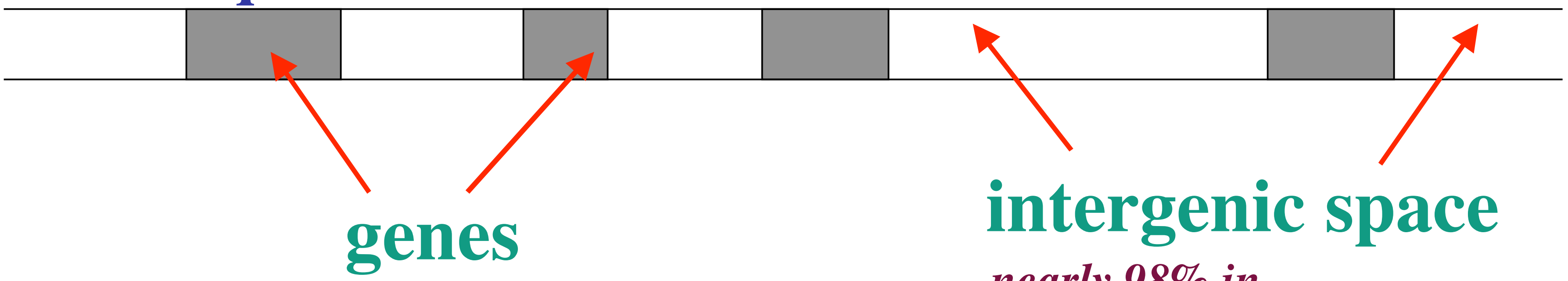
HMM introns
exons
CpG islands
intergenic

Noan Chomsky, 1928-- computational linguist, MIT

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

DNA sequence



nearly 98% in
higher mammals like us!
Often called junk DNA

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

Recall Crick’s Central dogma of molecular biology:

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

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!

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

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).

Today nc-RNA genes are recognized to be extemely crucial to the functioning of cells.

Heriditary information is carried by

1. Protein-coding genes (known for many years).
2. ncRNA genes.
3. Epigenetic layers

Scientific American, December 2003

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).

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

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.

xx**AATAGCGA**xxxxxxxxxxxx**AATAC**xxx**AAATACCG**

xxxxxxAATAGCGAxxxxxAATACxxxxxAAATACCG

xxxxxxAAGAGCGAxxxxxAATACxxxxxAAAGTCCG

xxxxxxAAAGCGAxxxxxAATACxxxxxAAATAAACCG

Multiple-alignment problem with gaps and mutations

Scoring problem

Hidden Markov models, again useful. ***Lots of good problems for theoreticians!***

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)

Examples

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

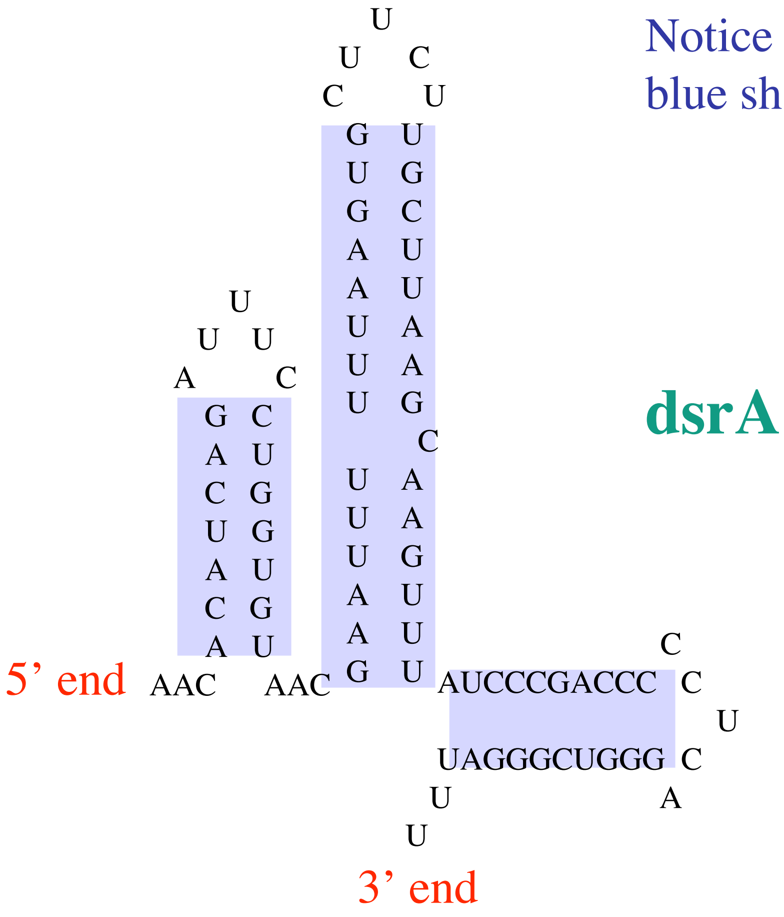
Intergenic space = biological dark matter?

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!

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

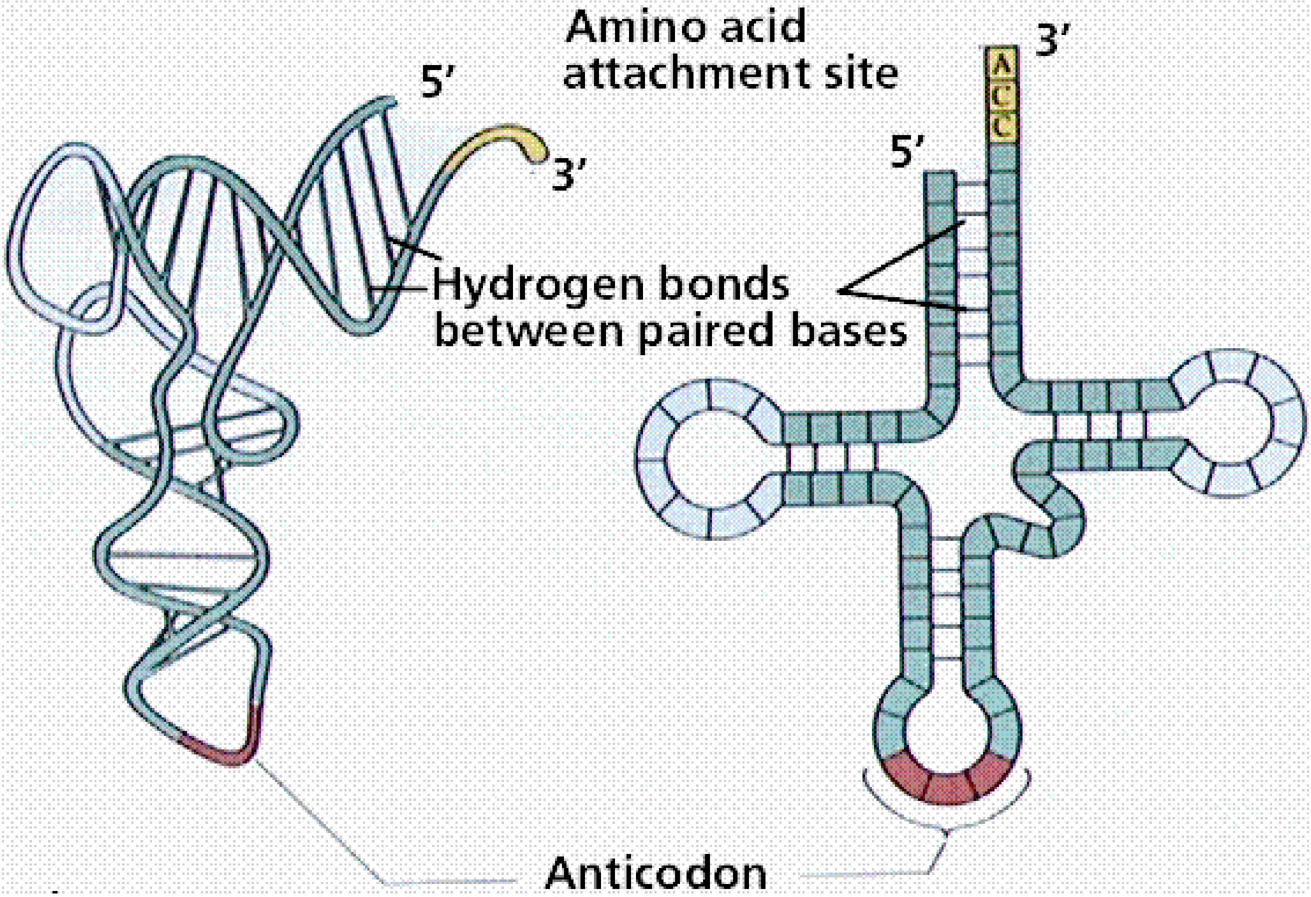
See S. R. Eddy, Nature Reviews, 12/01
for many examples and detailed discussions

A ═ U

C ≡ G

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

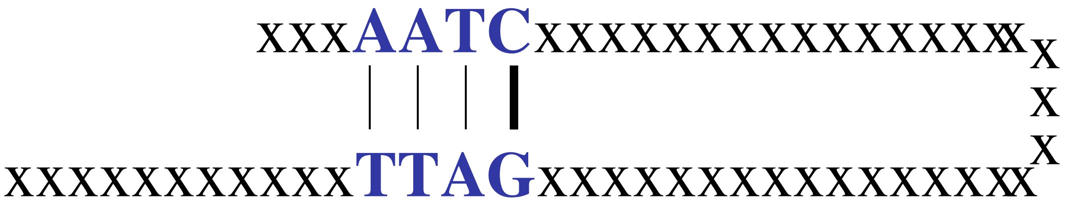
tRNA molecule (clover-leaf form)



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

Notice amazing amount of secondary structure xxx**AATC**xxxxxxxxxxxxxxxxxxxxxxxx**GATT**xxxxxxxxxxx

Linear sequence representing an ncRNA-gene

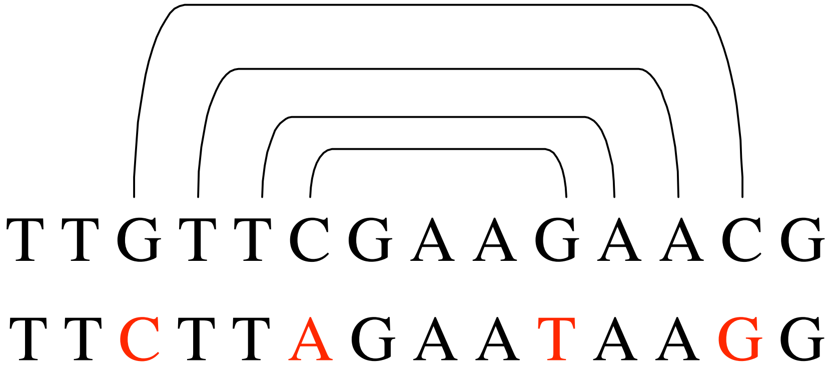


Folded sequence

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.



T T G T T C G A A G A A C G

T T C T T A G A A T A A G G

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.

A = T

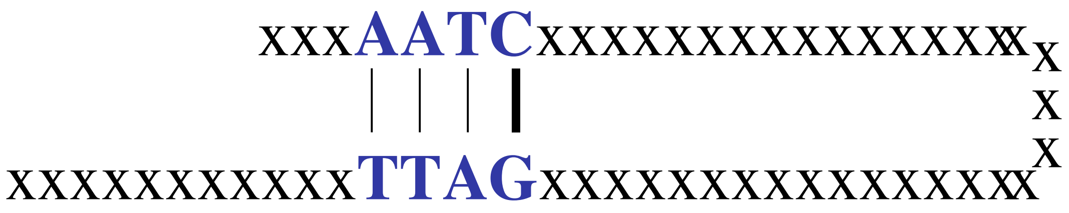
C ≡ G

Routine steps in the application of HMM

Given the HMM and an output sequence y(1),y(2), ….
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), ….
how to compute the probability that the HMM generates this?
Forward-backward algorithm

How to adjust the model parameters □ and □ such that they are optimal for an application, e.g., to represent exons?
Training; Expectation Maximization algorithm (Baum-Welch).



xxxAATCxxxxxxxxxxxxxxxxx
x
x
x
xxxxxxxxxxxxTTAGxxxxxxxxxxxxxxxxx

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,Vancover

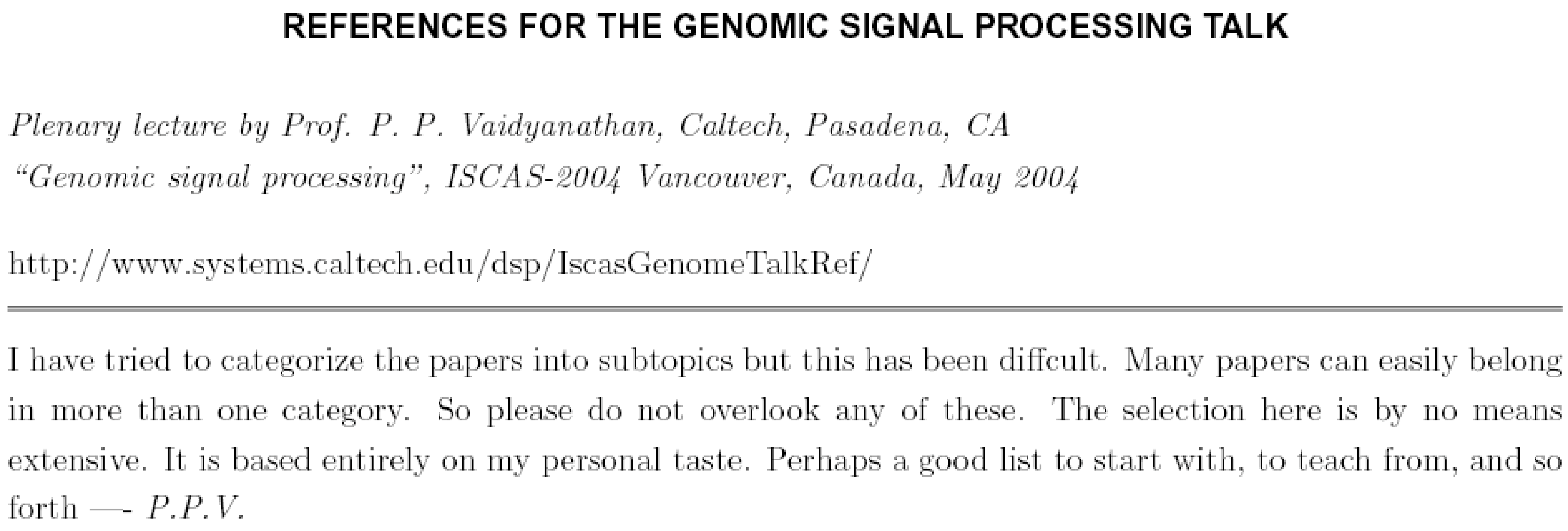
Biology today is not just wet stuff in smelly labs!

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

End of this part

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



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/IscasGenomeTalkRef/

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.*

Continue →

The great paper

The paper which started it all ...

[1] J. D. Watson, and F. H. C. Crick, A structure for DNA, Nature, April 1953.

Books and Tutorials

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*, Garland Publishing Inc., New York, 1998.

[2] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, pp. 8–20, July 2001.

[3] P. P. Vaidyanathan, and B-J. Yoon, "The role of signal processing concepts in genomics and proteomics," Journal of the Franklin Institute, vol. 341, pp. 111–135, 2004.

[4] Proc. of IEEE special issues Dec. 2000 (Genomic Engineering), Nov. 2002 (Bioinformatics, part 1: advances and challenges), and Dec. 2002 (Bioinformatics, part 2: genomics and proteomics engineering).

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[6] S. L. Salzberg, D. B. Searls, and S. Kasif, *Computational methods in molecular biology*, Elsevier, 1998.

[7] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, 1998.

Signal-processing flavor (DNA/Protein)

[1] I. Cosic, “Macromolecular bioactivity: is it resonant interaction between macromolecules? — theory and applications”, IEEE. Trans. Biomedical Engr., vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

[2] W. Huang, D. R. Fuhrmann, D. G. Politte, L. J. Thomas, and D. J. States, “Filter matrix estimation in automated DNA sequencing,” IEEE Trans. on Biomedical Engr., vol. 45, no. 4, pp. 422–428, April 1998.

[3] S. W. Davies, M. Eizenman, and S. Pasupathy, “Optimal structure for automatic processing of DNA sequences,” IEEE Trans. on Biomedical Engr., vol. 46, no. 9, pp. 1044–1056, Sept. 1999.

[4] X-P. Zhang, and D. Allison, “Iterative deconvolution for automatic base scaling of the DNA electrophoresis time series,” Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.

[5] E. Pirogova, Q. Fang, M. Akay, and I. Cosic, “Investigation of the structural and functional relationships of oncogene proteins”, Proc. of the IEEE, vol. 90, no. 12, pp. 1859–1867, Dec. 2002.

[6] D. Sussillo, A. Kundaje, and D. Anastassiou, “Spectrogram analysis of genomes”, Eurasip J. of Applied Signal Processing, vol. 2003, no. 4, Dec. 2003.

[7] M. L. Simpson, C. D. Cox, G. D. Peterson, and Gary S. Sayler, “Engineering in the biological substrate: information processing in genetic circuits,” Proc. of the IEEE, vol. 92, no. 5, pp. 848–863, May 2004.

Gene prediction

[1] A. Krogh, I. Saira Mian, and D. Haussler, “A hidden Markov model that finds genes in E. Coli DNA”, Nucleic Acids Research, vol. 22 pp. 4768–4778, 1994.

[2] J. W. Fickett, “The gene prediction problem: an overview for developers”, Computers Chem., vol. 20, no. 1, pp. 103–118. 1996.

[3] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, “Prediction of probable genes by Fourier analysis of genomic sequences,” CABIOS, vol. 13, no. 3, pp. 263–270, 1997.

[4] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, “Microbial gene identification using interpolated Markov models,” Nucleic Acids Research, vol. 26, no. 2, pp. 544–548, 1998.

[5] P. P. Vaidyanathan, and B-J. Yoon, “Gene and exon prediction using allpass-based filters,” Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.

[6] P. P. Vaidyanathan, and B-J. Yoon, “Digital filters for gene prediction applications,” IEEE Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2002.

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

[1] E. N. Trifonov, and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence", Proc. of the Nat. Acad. Sci., USA, vol. 77, pp. 3816–3820, 1980.

[2] G. W. Wornell, "A Karhunen-Loeve-like expansion for $1/f$ processes via wavelets," IEEE Trans. on Information Theory, vol. 36, no. 4, pp. 859–861, July 1990.

[3] W. Li, "Expansion-modification systems: A model for spatial $1/f$ spectra", Physical review A, The American Physical Society, vol. 43, no. 10, pp. 5240–5260, May, 1991.

[4] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," Nature, vol. 356, pp. 168–170, March 1992.

[5] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," Physical Review Letters, vol. 68, no. 25, pp. 3805–3808, June 1992.

[6] H. Hausdorff and C.-K. Peng, "Multiscaled randomness: a possible source of $1/f$ noise in biology," Physical review E, vol. 54, no, 2, pp. 2154–2157, August 1996.

[7] W. Li, "The study of correlation structures of DNA sequences: a critical review", Computers Chem., vol. 21, no. 4, pp. 257–271, 1997.

[8] H. Herzel, E. N. Trifonov, O. Weiss, and I. Groβe, "Interpreting correlations in biosequences," Physica A, vol. 249, pp. 449–459, 1998.

[9] M. de Sousa Vieira, “Statistics of DNA sequences: a low-frequency analysis,” Physical Review E, The American Physical Society, vol. 60, no. 5, pp. 5932–5937, Nov. 1999.

[10] Z-G. Yu, V. V. Anh, and B. Wang, “Correlation property of length sequences based on global structure of the complete genome”, Physical review E, The American Physical Society, vol. 63, pp. 011903-1—011903-8, 2000.

[11] K. B. Murray, D. Gorse, and J. M. Thornton, “Wavelet transforms for the characterization and detection of repeating motifs,” J. Molecular Biology, vol. 316, pp. 341–363, 2002.

Noncoding RNA, Noncoding genes

[1] S. R. Eddy, "Noncoding RNA genes and the modern RNA world," Nature reviews, GENETICS, vol. 2, pp. 919–929 Dec. 2001.

[2] S. R. Eddy, "Computational genomics of noncoding RNA genes," Cell, vol. 109, pp. 137–140, April 2002.

[3] G. Storz, "An expanding universe of noncoding RNAs," Science, vol. 296, pp. 1260–1263, May 2002.

[4] N. C. Lau and D. P. Bartel, *Censors of the genome*, Scientific American, pp. 35-41, August 2003.

[5] W. W. Gibbs, *The unseen Genome: gems among junk*, Scientific American, pp. 48-53, Nov. 2003.

[6] W. W. Gibbs, *The unseen Genome: beyond DNA*, Scientific American, pp. 108-113, Dec. 2003.

[7] W. W. Gibbs, *Synthetic life*, Scientific American, pp. 75-81, May 2004.

DNA microarrays

[1] P. O. Brown and D. Botstein, “Exploring the new world of the genome with DNA microarrays,” Nature genetics supplement, vol. 21, pp. 33-37, Jan. 1999.

[2] E. S. Lander, “Array of hope,” Nature genetics supplement, vol. 21, pp. 3-4 Jan. 1999.

[3] C. Debouck and P. N. Goodfellow, “DNA microarrays in drug discovery and development,” Nature genetics supplement, vol. 21, pp. 48–50, Jan. 1999.

[4] O. Alter, P. O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling”, Proc. of the Natl. Acad. of Sci., USA, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.

[5] S. K. Moore, “Making chips to probe genes”, pp. 54–60, IEEE Spectrum, vol. 38, no. 3, March 2001.

[6] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, “Iterative normalization of cDNA microarray data,” IEEE Trans. on Information Tech. in Biomedicine, vol. 6, no. 1, pp. 29–37, March 2002.

[7] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, “Iterative normalization of cDNA microarray data,” IEEE Trans. Info. Tech. in Biomed., vol. 6, no. 1, pp. 29–37, March 2002.

[8] R. Casagrande, *Technology against terror*, Scientific American, pp. 82-87, Oct. 2002.

[9] O. Alter, P. O. Brown, and D. Botstein, “Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms,” Proc. of the Natl. Acad. of Sci., USA, vol. 100, no. 6, pp. 3351–3356, March 2003.