Signal Processing in Bioinformatics

slide set # 6

Alessandro Falaschi

Signal Processing and Information Theory
Bioinformatics Degree, Sapienza University of Roma

Didactic material on TeoriadeiSegnali.it

April 2023

Overwiev

Summary of the presentation

We address here several bioinformatics topics from the point of view of signal processing. First, a method to identify the genes responsible for cell division.

Then the frequency analysis of the genome, aimed at detecting DNA coding regions within ORFs by FFT analysis, due to the repeated presence of codons with length 3. The sequence coding of the genome can be based on indicators, or on more compact encodings either based on tetrahedra or complex mapping. Some encodings have also been proposed for the codons sequence. Finally, an evolutionary explanation is provided for the long-range correlation of the genome, which corresponds to a slope $\frac{1}{f}$ of its overall spectrum

A further topic concerns the spectral analysis of the amino acid sequence of proteins, carried out on the basis of their EIIP. This allows a functional role to be attributed to proteins based on a characteristic frequency revealed by the consensus spectrum method

Finally, the Fourier transform infrared spectroscopy method is analysed, in which an interferometer is studied as if it were a filter with one tap, and the variation of the delay, together with Wiener's theorem, allows the absorption spectrum of a substance to be measured, and to infer its molecular composition.

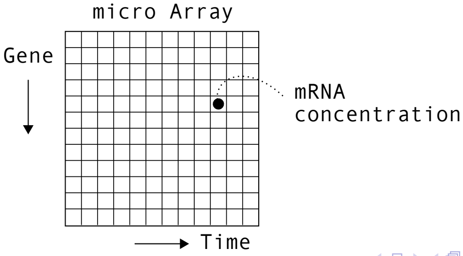
Content index

1. **The cell cycle is periodic**
2. **The genome is a sequence: Genome frequency analysis**
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. **Fourier transform infrared spectroscopy (FTIR)**
   - A tool for identifying chemical substances or functional groups

Finding the genes that drive mitosis

The cell cycle is periodic

- A cell divides into two new cells by following a sequence of phases called G1, S, G2 and M
  - sequence flow is driven by the formation of CDK-cyclin complexes, as discovered in 1995
  - the concentration of cyclin proteins varies as a consequence of the expression level of some genes, the **cdc** (*cell division cycle*) genes
- The identities of **cdc** genes were first determined in 1998 by microarray hybridisation, in order to analyze the mRNA levels in cell cultures that had been synchronized



The Fourier score

- mRNA concentrations were used for the computation of a so-called *Fourier Score* as a measure of the periodicity of gene expression
  - apart from a series of normalizations, this quantity has been defined as

    $$A = \sum_{n=0}^{39} e_n \sin \frac{2\pi}{40} n$$

    $$B = \sum_{n=0}^{39} e_n \cos \frac{2\pi}{40} n$$

    $$Fourier\ score = \sqrt{A^2 + B^2}$$

    where e_n is the sequence of gene expression levels based on the mRNA concentration levels

- Evidently here we are evaluating the energy of the first DFT coefficient for the expression profile of the gene that transcribes that mRNA
  - in this sense, it detects whether the expression profile makes *a complete change* during the time interval of a cell's division cycle

Now we talk about...

1 The cell cycle is periodic

2 **The genome is a sequence: Genome frequency analysis**

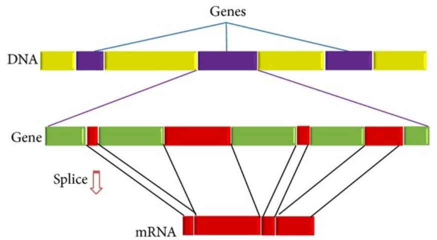
- Biochemical information encoding
- Spectral analysis of the genome
- Filtering of the genome
- Other representation spaces
- Numerical representation of codons
- Long range DNA correlation
- Fourier transform of protein sequences

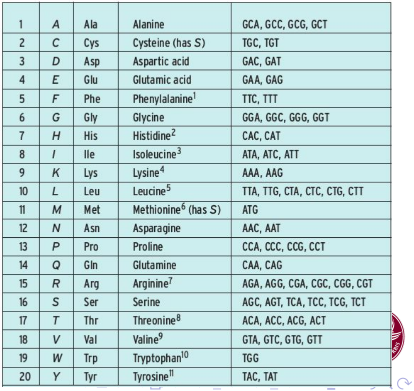
3 Fourier transform infrared spectroscopy (FTIR)

- A tool for identifying chemical substances or functional groups

DNA is an *interleaved* message

- Nearly 98% of DNA in higher mammals is made of *intergenic space* (yellow): it does not encode proteins, but nonetheless translate into ncRNA with functions *inside* the cell
- *Genes* (purple) are transcribed into *mRNA* which in turn translates into proteins, but only after *introns* (green) have been spliced off from *exons* (red)
- Each group of three *bases* (A, T, C, G) within the exons ($4^3 = 64$ possibilities) was called a *codon*, and will translate to one of the 20 different ammino acid that *proteins* are made of
- Codons have a very *uneven* frequency distribution, see [1] and [2]
- Nucleotides tend to appear more often in the same position within codons belonging to the same exon, while in introns their position is much more *random*
- **Result**: the coding regions of DNA are characterized by a strong periodic component!





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

Now we talk about...

1. The cell cycle is periodic
2. The genome is a sequence: Genome frequency analysis
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. Fourier transform infrared spectroscopy (FTIR)
   - A tool for identifying chemical substances or functional groups

DNA numerization

The periodic repetitiveness of the codon composition can be exploited to identify the coding regions of the DNA

- there are actually *two strands* of DNA, and for each of them there are three different ways of aligning the sequence of bases (called open reading frame, i.e. a sort of windowing of the genome) to a supposed sequence of codons

```
5′- CATTGCCAGT - 3′
3′- GTAACGGTCA - 5′

   CAT TGC CAG T..
   .CA TTG CCA GT.
   ..C ATT GCC AGT
   ACT GGC AAT G..
   .AC TGG CAA TG.
   ..A CTG GCA ATG
```

After the DNA has been sequenced, only a *symbolic* string is available: The letters A, T, C, G must be replaced by numbers, thus obtaining a *numerical* sequence, that can be processed by FFT.

- Voss introduced the use of four *indicator sequences* $u_n^a, u_n^t, u_n^g, u_n^c$, one for each base, with value 1 or zero for indices n in which the relative base is present or not

DNA	A	T	T	G	C	A	C	C	G	T	G	A
u_n^a	1	0	0	0	0	1	0	0	0	0	0	1
u_n^t	0	1	1	0	0	0	0	0	0	1	0	0
u_n^g	0	0	0	1	0	0	0	0	1	0	1	0
u_n^c	0	0	0	0	1	0	1	1	0	0	0	0

- Richard F. Voss, *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences*, Phys. Rev. Lett. 68, 3805 – 22 June 1992 - https://doi.org/10.1103/PhysRevLett.68.3805

DNA spectral analysis

From each indicator sequence a DFT can be calculated on a window of length N *multiple of three*, thus obtaining *four sequences* of spectral coefficients

$$U_m^a \qquad U_m^t \qquad U_m^g \qquad U_m^c$$

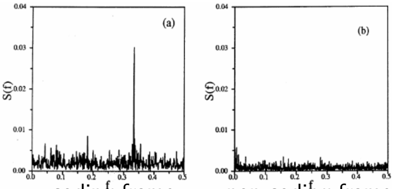
each relating to the periodicities observed by the associated base, within the analysis window

Example: $$U_m^a = \sum_{n=0}^{N-1} u_n^a e^{-j2\pi \frac{m}{N} n}$$

To highlight the periodicity information from all four bases, a kind of *power spectral density* is defined, as

$$S_m = (U_m^a)^2 + (U_m^t)^2 + (U_m^g)^2 + (U_m^c)^2$$

When the window is located in a *coding region* the powed density S_m presents a pronounced peak at the *digital frequency* of $\frac{2}{3}\pi$, which is instead absent for windows taken inside of non-coding regions



coding frame non coding frame

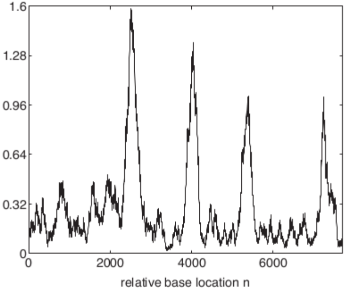
To span an integer number of codons the window length should be a multiple of three, such as N =351 or larger, so that the periodicity effect dominates the background ${}^{1}\!/_{f}$ spectrum (page 6), but not too long, so as not to compromise the *spatial resolution*

DFT indices *vs.* sequence periodicity

- Recall that S_m with $m = 0, 1, \cdots, N-1$ identifies the spectral component of frequency $f = \frac{m}{N} f_c$
  - but the genome does not have sampling frequency (at most, an interbase distance of few Å), so we are interested in expressing the information on the frequency in terms of period, i.e. after how many bases there is a repetitive behaviour
  - further, we reflect that a peak at $\frac{2}{3}\pi$ of the DTFT corresponds to a peak at $\frac{2}{3}\frac{f_c}{2} = \frac{f_c}{3}$
    - ★ but f_c maps to N in the DFT indicies, so that $\frac{f_c}{3}$ maps to $\frac{N}{3}$
- more generally, the frequency content at index m of a DFT refers to a periodicity p of $p = \frac{N}{m}$ samples
  - so the peak at $\frac{2}{3}\pi$, i.e. to index $\frac{N}{3}$, refers to a period $p = \frac{N}{m} = \frac{N}{N/3} = 3$ samples, or three bases - which is the length of a codon
  - this relation makes sense up to $m \leq \frac{N}{2}$, i.e. up to the first half of the indices, corresponding to a minimum periodicity $\frac{N}{m}$ of two samples, because less would violate the minimum of two samples per period

Plotting of the N/3 value from the DFT

- Given the DFT power spectrum S_m of a window of a sequenced DNA strand, the amplitude of its element $S_{N/3}$ could indicate whether the window belongs to a coding region



- We now denote by $S_{N/3}(n)$ the sequence of $S_{N/3}$ values collected as the window moves along the DNA chain, with the index n representing the base count from the start of the ORF

- $S_{N/3}(n)$ can be plotted to represent how the indicator of a coding region varies along the strand

- but remember that the values of a DFT can be interpreted as the outputs of a filter bank, and $S_{N/3}(n)$ as the output of the $N/3-$th filter of the bank, which corresponds to a sinc-type frequency response
  - therefore, the output values also depend on the signal components found at other frequencies of the short-time spectrum of the genome

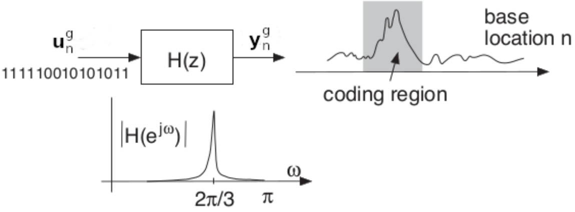
- a more accurare graph can be otained by means of an ad-hoc bandpass filter

Now we talk about...

1. **The cell cycle is periodic**
2. **The genome is a sequence: Genome frequency analysis**
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. **Fourier transform infrared spectroscopy (FTIR)**
   - A tool for identifying chemical substances or functional groups

Filters which have been proposed

The picture on the rigth shows the filtering of the Voss indicator sequence u_n^g through a narrow bandpass filter, and the assumed output y_n^g

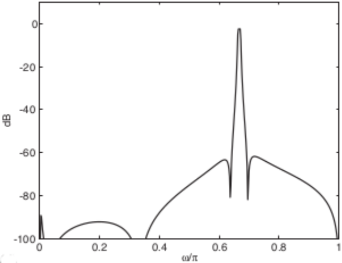


The experiment was actually conducted using a third order *antinotch* IIR filter whose frequency response can be adjusted in order to find a good compromise between sufficiently narrow bandwidth and acceptable spatial resolution

- but the outcome is still rather noisy along the non-coding regions

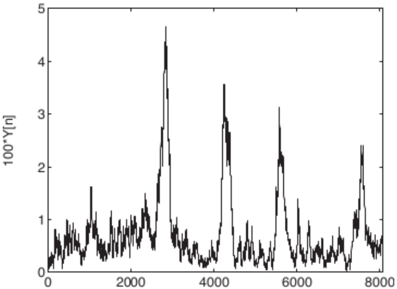
Another design has led to the implementation of a *multi-stage* (IFIR) filter with much better stopband attenuation, as can be seen in the figure, at the price of only slightly higher complexity

In this case the non-coding regions behave more smoothly

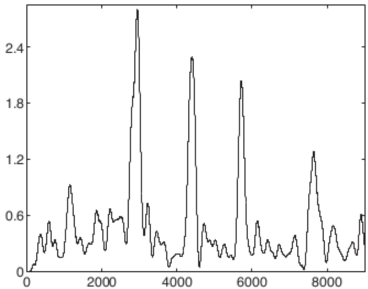


Smoothed exon prediction results

The output signal from each of the two filters described is shown below, as a function of the base number



antinotch



multistage

- Full details are given in Vaidyanathan, P. P., & Yoon, B. J. (2004). The role of signal-processing concepts in genomics and proteomics. Journal of the Franklin Institute, 341(1-2)

- Further techniques can be found in Tuqan, J., & Rushdi, A. (2008). A DSP approach for finding the codon bias in DNA sequences. IEEE Journal of Selected Topics in Signal Processing, 2(3), 343-356

Now we talk about...

1. The cell cycle is periodic
2. The genome is a sequence: Genome frequency analysis
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. Fourier transform infrared spectroscopy (FTIR)
   - A tool for identifying chemical substances or functional groups

Tetrahedron and the 3D space of bases

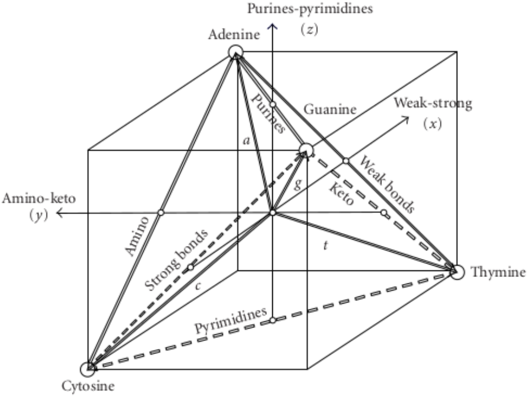
First we observe that $u_n^a + u_n^t + u_n^g + u_n^c = 1$ for each index n, meaning that any indicator can be obtained from the other three, i.e. that it is possible to resort to a representation of reduced dimensionality without losing information

One way to do this is to place *the bases* ATCG on the vertices of a regular tetrahedron, which form a subset of the vertices of a *cube* with side lengths equal to two. Their position is chosen according to their chemical-physical properties

In this way the nucleotide sequence can be associated with a 3D coeffient vector along the x, y, z axis as

	A	T	C	G
x	1	1	-1	-1
y	1	-1	1	-1
z	1	-1	-1	1

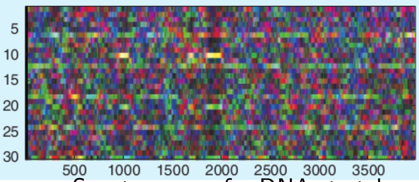
thus preserving the differences “weak - strong’ bonds, “amino - keto”, and “purines - pyrimidines”



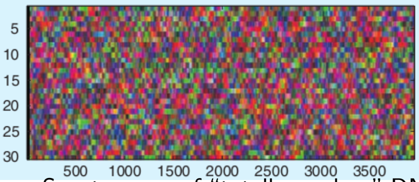
DNA spectrograms

The same DFT processing done for the indicator sequences can be applied to the new 3D x_n, y_n, z_n representaton, obtaining three X_m, Y_m, Z_m spectra, and a power spectral density, obtaining results comparable with the previous case

Here is the spectrogram It arises after interpreting X_m, Y_m, Z_m as the values of a red, green and blue (RGB) color component for frequency m, and repeating the spectral analysis on successive windows along the DNA strand



Spectrogram of a DNA stretch

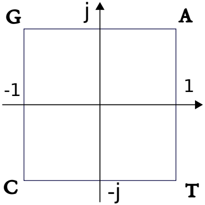


Spectrogram of "totally random" DNA

- The spectrogram shows DFTs of *length* $N = 60$ of a stretch of DNA of 4,000 bases: vertically the k frequencies from 1 to 30 (half the length), and horizontally the positions of the nucleotides
- There are three coding regions at 953-1066, 1668-1727 and 1807-2028, corresponding to the brightest bars at frequency $k = 10$ (period 6 corresponds to $^N/_6 = 10$). But we also see the presence of other frequencies
- Right, a totally random DNA spectrogram, with the probability of bases of uniform density, memoryless
- more on this on Sussillo Kundaje Anastassiou, Spectrogram Analysis of Genomes, EURASIP J on App. SP 2004, Hindawi Publishing Corporation

DNA bi-dimensional projection

The dimensionality of the tetrahedral representation can be reduced to two by projecting it onto a plane: the simplest choice is a projection from above



We can then map the coordinates as a complex number

$$\begin{aligned} A &= 1 + j \\ T &= 1 - j \\ C &= -1 - j \\ G &= -1 + j \end{aligned}$$

In this way the complementarity of the base pairs $A - T$ and $C - G$, respectively, is expressed by the conjugate symmetry $T = A^*$, $G = C^*$ making it possible to express the complementary strand as *the conjugate* of the numerical coding for the reading frame in analysis.

A comparison of this and other numerical mappings of the genome can be found at

- Kwan, H. K., & Arniker, S. B. Numerical representation of DNA sequences, in 2009 IEEE International Conference on Electro/Information Technology (pp. 307-310), IEEE
- Yu, N., Li, Z., & Yu, Z. (2018) Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning Big Data Mining and Analytics, 1(3), 191-210

Now we talk about...

1. The cell cycle is periodic
2. The genome is a sequence: Genome frequency analysis
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. Fourier transform infrared spectroscopy (FTIR)
   - A tool for identifying chemical substances or functional groups

3-D Codon representation

Instead of a four-base alphabet, codons have a 64-letter alphabet, for which a tetrahedral model and two-dimensional numerical encoding can be derived

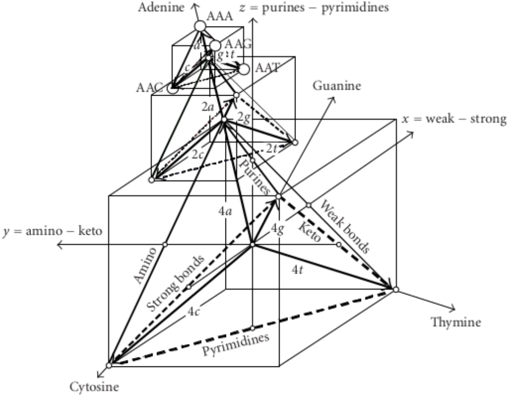
This is possible by repeating the graphic construction three times, as many as there are bases inside a codon

Each time a tetrahedron is added, its width is reduced by half, in a process similar to that of the base two expression of a number

Correspondingly, the 3-D vector for codon X can be built as

$$\overline{x} = 2^2\overline{b}_2 + 2^1\overline{b}_1 + 2^0\overline{b}_0$$

with $\overline{b}_i \in \{\overline{a}, \overline{c}, \overline{g}, \overline{t}\}$; $i = 0, 1, 2$



and $\overline{a}, \overline{c}, \overline{g}, \overline{t}$ are the coordinate vectors of the position of the bases in the tetrahedron

More on this in the first chapter of Dougherty, E. R., & Shmulevich, I. (Eds.). (2005). Genomic signal processing and statistics (Vol. 2). Hindawi Publishing Corporation

2-D Codon representation

A similar process can be performed by entering the bi-dimensional complex encoding of bases into a FIR filter

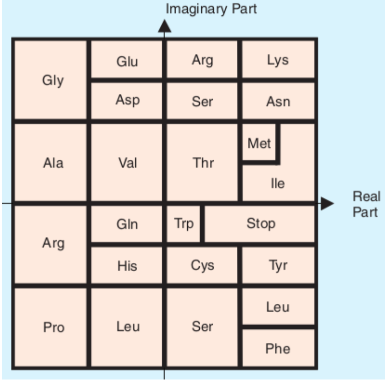
$y[n] = h[0]x[n] + h[1]x[n-1] + h[2]x[n-2]$

with *taps* $h[0] = 1$, $h[1] = 1/2$, and $h[2] = 1/4$.

Then the elements of the output subsequence

$$y[2], y[5], y[8], y[11], ..., y[N-1]$$

are complex numbers representing each of the amino acids of the resulting protein



In fact, the entire genetic code can be drawn on the complex plane as shown in Fig, in which the Met*ionine* (coded by ATG) is the complex number

$$(1+j) + 0.5 \cdot (1 – j) + 0.25 \cdot (-1+j) = 1.17 + 0.88 \cdot j$$

Each of the entries in Fig. correspond to one of the 20 amino acids or the STOP codon. Therefore, the protein coding process can be simulated by a digital low-pass filter, followed by subsampling via a three-band polyphase decomposition, followed by a switch selecting one of the three bands (reading frames), followed by a vector quantizer as defined in Fig.

Now we talk about...

1. The cell cycle is periodic
2. The genome is a sequence: Genome frequency analysis
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences
3. Fourier transform infrared spectroscopy (FTIR)
   - A tool for identifying chemical substances or functional groups

Long range DNA correlation (or 1/f behavior)

DNA sequences exhibit a long-range correlation also, both in the gene and intergenetic regions, over much longer regions which contained many genes

- according to WIENER theorem, long range correlation implies that the Fourier transform has a strong low-frequency content
- the *autocorrelation* can be calculated starting from an indicator sequence (let's say, the one of Adenine) as

$$\mathcal{R}_A(k) = \sum_n u^a(n)\, u^a(n+k)$$

and from it (by Wiener theorem)

$$\mathcal{P}_A\left(e^{j\omega}\right) = DTFT\left\{R_A(k)\right\}$$

The lowest frequency measured by a DFT is $\frac{1}{N}f_c$, or $\frac{1}{N}2\pi$, or $\frac{1}{N}$ if we normalize $f_c = 1$, and Voss demonstrated that $\mathcal{P}_A\left(e^{j\omega}\right)$ obey to a power-law ${}^1/_{f^\beta}$, $\beta \simeq 1$, for each of the four indicator sequences; later studies found that such correlation extends to several millions of bases

Long range DNA correlation (or 1/f behavior)

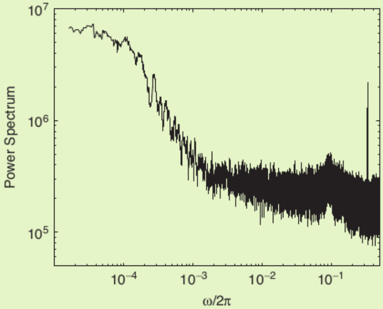
In figure it is shown the $\mathcal{P}_A\left(e^{j\omega}\right)$ spectrum (in a log-log plot) for base A for the first one million bases of a bacterial genome, so that there were 0.5 million samples of $\mathcal{P}_A\left(e^{j\omega}\right)$ in $0 \leq \omega \leq \pi$

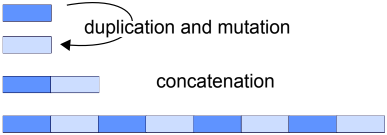
The variations near zero frequency can be seen clearly, and the $1/f$ behavior continues till very low frequencies, flattening out only as we get really close to zero frequency

The thin line near the right edge of the plot corresponds to the peak at $2\pi/3$ due to period-3 property in the coding regions

The $1/f$ behavior can be traced to the so-called duplication-mutation model, see figure

It has been observed that DNA molecules also have components of period 10 to 11




duplication and mutation

concatenation

long range correlation

It is argued that this periodicity can be attributed to an alternation property in protein molecules. This arises from the fact that the hydrophilic and hydrophobic regions alternate at a certain rate in the three-dimensional folded form

Now we talk about...

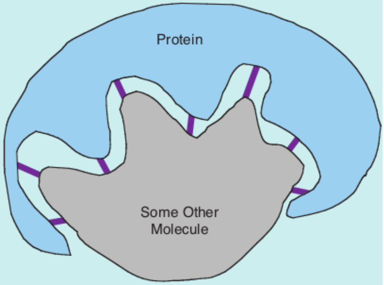
1. The cell cycle is periodic

2. **The genome is a sequence: Genome frequency analysis**
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences

3. Fourier transform infrared spectroscopy (FTIR)
   - A tool for identifying chemical substances or functional groups

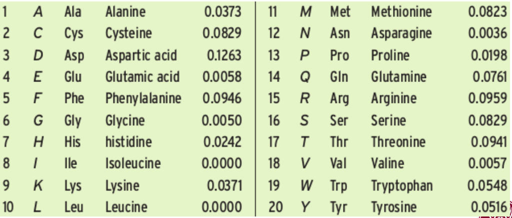
Proteins as sequences of EIIP

Proteins do interact selectively with other molecules thanks to their 3D shape. There are specific sites in the proteins 3D structure called *hot spots* where certain other molecules can conveniently bind to the protein



With each one of the twenty amino acids it is possible to associate a unique nonnegative number called the average *electron-ion interaction potential* (EIIP)

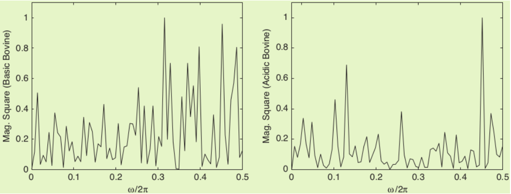
These EIIP values are natural candidates for the numerical representation of the amino acid sequence associated with a protein, on which a DFT can be calculated



1	*A*	Ala	Alanine	0.0373	11	*M*	Met	Methionine	0.0823
2	*C*	Cys	Cysteine	0.0829	12	*N*	Asn	Asparagine	0.0036
3	*D*	Asp	Aspartic acid	0.1263	13	*P*	Pro	Proline	0.0198
4	*E*	Glu	Glutamic acid	0.0058	14	*Q*	Gln	Glutamine	0.0761
5	*F*	Phe	Phenylalanine	0.0946	15	*R*	Arg	Arginine	0.0959
6	*G*	Gly	Glycine	0.0050	16	*S*	Ser	Serine	0.0829
7	*H*	His	histidine	0.0242	17	*T*	Thr	Threonine	0.0941
8	*I*	Ile	Isoleucine	0.0000	18	*V*	Val	Valine	0.0057
9	*K*	Lys	Lysine	0.0371	19	*W*	Trp	Tryptophan	0.0548
10	*L*	Leu	Leucine	0.0000	20	*Y*	Tyr	Tyrosine	0.0516

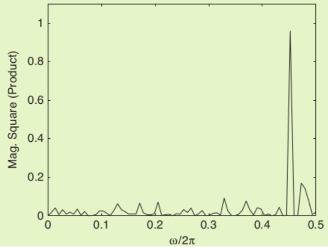
Proteins transform and their consensus spectrum

Let X_m be the values of the DFT of the EIIP sequence obtained for one protein, and Y_m the values obtained for another



Although the graph of their modules $|X_m|$, $|Y_m|$ does not reveal anything of interest, if the two proteins have some function *in common* (i.e. they bind to the same kind of molecule) their spectra contain a peak at a frequency in common for both proteins, which can be revealed by multiplying the two spectra, i.e. evaluating $|X_m| \cdot |Y_m|$ which is also called the *consensus spectrum*

If a protein performs more than one function, each function correspond to a unique *characteristic frequency*



Sahoo, S.S., & Hota, M.K. (2014). Determination of Characteristic Frequency for identification of Hot spots in Proteins using Computational Simulations: a Review. American Journal of Computing Research Repository, 2(2), 38-43

Further reading on genomic DSP

- P. Ramachandran, A. Antoniou, Genomic Digital Signal Processing (slides)
- D. Anastassiou, Genomic Signal Processing, IEEE Signal Processing Magazine, 2001
- P.P. Vaidyanathan, Genomics and Proteomics: a Signal Processor's Tour (2004) IEEE Circuits and Systems Magazione
- P.P.Vaidyanathan, Byung-Jun Yoon, The role of signal-processing concepts in genomics and proteomics, Journal of the Franklin Institute, Volume 341, 2004, https://doi.org/10.1016/j.jfranklin.2003.12.001
- H. K. Kwan and S. B. Arniker, Numerical representation of DNA sequences, 2009 IEEE International Conference on Electro/Information Technology, 2009, pp. 307-310, doi: 10.1109/EIT.2009.5189632
- Xin-Yun Zhang et al., Signal processing techniques in genomic engineering, in Proceedings of the IEEE, vol. 90, no. 12, pp. 1822-1833, Dec. 2002, doi: 10.1109/JPROC.2002.805308
- Sussillo, D., Kundaje, A., Anastassiou, D. Spectrogram Analysis of Genomes. EURASIP J. Adv. Signal Process. 2004, 790248 (2004). https://doi.org/10.1155/S1110865704310048
- Those found in this Google Drive (but I did not ask for permission to editors)

Now we talk about...

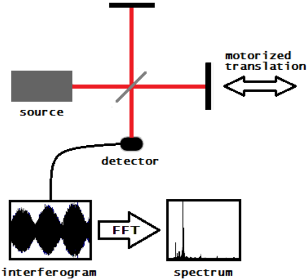
1. **The cell cycle is periodic**

2. **The genome is a sequence: Genome frequency analysis**
   - Biochemical information encoding
   - Spectral analysis of the genome
   - Filtering of the genome
   - Other representation spaces
   - Numerical representation of codons
   - Long range DNA correlation
   - Fourier transform of protein sequences

3. **Fourier transform infrared spectroscopy (FTIR)**
   - A tool for identifying chemical substances or functional groups

Infrared spectroscopy

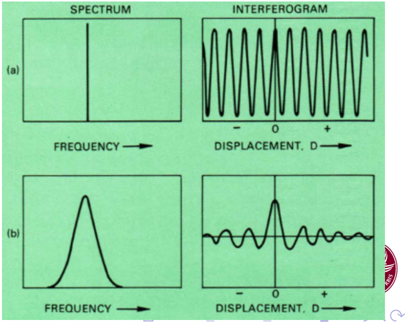
- Spectroscopy is a means of determining the atomic composition of a material based on the wavelengths λ *of the light* it absorbs, or doesn't let through. Tables: [1] [2]
  - it can be done by measuring one λ at a time (*monochromatic light*), or
  - by using all the λ at the same time, by varying their mix as a function of time, and then apply a Fourier transform
- The second method uses an interferometer, in which light from a broadband source is split by a semi-reflective mirror to two other mirrors, and then combined back and passed *through the sample* to be measured
- The light source has a power density $\mathcal{P}(f)$, and the detector measures the its *intensity* $I = \int \mathcal{P}(f)\, df$
- The ΔL difference in path length causes each λ to be attenuated or boosted. Video: [1] [2]



- By moving one of the two reflecting mirrors the mix of the λ changes. At the end an *interferogram* is recorded, which depends on ΔL instead of time

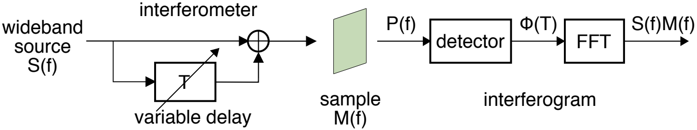
Fourier transform spectroscopy

- We will study that the sum of a signal with its delayed (by T seconds) copy *is equivalent* to its passage through a *comb* filter with frequency response $|H(f)|^2 = 1 + \cos 2\pi fT$, and that the power spectrum at the output end is equal to $\mathcal{P}(f)|H(f)|^2$
- The detector measures the energy from *all the frequencies*, i.e.
  $$I(T) = \int \mathcal{P}(f)|H(f)|^2 df = \int \mathcal{P}(f)(1 + \cos 2\pi fT)\, df = \mathcal{P} + \Phi(T)$$
  where the term $\Phi(T) = \int \mathcal{P}(f)\cos 2\pi fTdf$ varies with T and is called the *interferogram*
- Knowing that $\mathcal{P}(f)$ is a *real even* function, we recognize that $\Phi(T)$ is the inverse Fourier of $\mathcal{P}(f)$, so that $\mathcal{P}(f)$ can be retrieved by transforming $\Phi(T)$, that is
  $$\mathcal{P}(f) = \int \Phi(T)\cos 2\pi fTdT$$



Infrared Fourier transform spectroscopy

- The sample of material under examination put in front of the detector adds another filtering block, so that the *observed* $\mathcal{P}(f)$ is equal to $S(f)\,M(f)$
  - in which $S(f)$ and $M(f)$ respectively are the power density of the light source, and the *absorption spectrum* of the sample material



- after evaluating $S(f)$ *without* the sample, a new measurement is taken with the sample in the middle, and $M(f) = {}^{P(f)}/_{S(f)}$ is obtained
- Atom size, bond length and bond strength vary in molecules and the absorption of infrared radiation at different wavelength provides useful information about the sample structure
- no two organic compounds have the same IR spectrum, so a compound can be identified by its absorption peaks

Further readings: [1], [2], [3], [4], [5]; videos: [1], [2], [3]