# Course Overview

## slide set # 1

Alessandro Falaschi

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

Didactic material on TeoriadeiSegnali.it

March 2024

# Introduction

## Objective of this presentation

After the illustration of the telematic tools of reference for the course,
we introduce ourselves to each other

A brief introduction to the nature of signals and their processing follows,
in order to provide an overview of the objects
we will be dealing with

In accordance with what I have proposed to myself,
some applicative aspects of signal processing and information theory
in the biological, medical, genetic, microscopic and structural fields
are then outlined, up to a mention of the most recent applications

But, really, don't be afraid of that, I just wanted to pique your curiosity
to study signal processing, at least to understand anything
you still don't understand about more advanced topics!

# Content index



# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Personal biography

## Until the beginning of the new century

- Born in Rome, 65 years old, 76 kg
- A degree on Electronic Engineering in '83 with 110 cum laude
- A PhD on speech recognition in '88 by Hidden Markov Models
  - four years doing a lot of research through scholarships and grants
  - but now speech recognition have progressed a lot!
- Hired as a researcher at University of Perugia in '92
- I developed a telephonic social (IperAudio) - w.o. cell phone or FB
- Back to Rome on '95, research shifted on OFDM transmission
- I have been webmaster and postmaster at my department
  - I did made a PHP web application for my Dpt. which looked better than Infostud
  - I thought that being able to freely express oneself electronically would be a great advance for humanity; now I agree that social, chat, email and VoD are making us asocial and confused

# Personal biography - 2

- from 2002 to 2006 I worked on Internet multimedia, VoIP, streaming
  - there was no FaceBook or Youtube and no Skype!
- In 2006-07 I worked on Vehicular Networking (VANET)
  - I liked the possibility of communicate in between cars without the need of a telephonic company
- from 2008 to 2019 I was asked to go to teach in Latina (at the Sapienza didactic pole)
  - where I taught three different telecommunication courses
  - covering 30.000 Km (also if for three years I have rented a room)
  - losing contact with my colleagues and halting the research
  - if I had known that there would be Covid I would have continued to teach at Latina in DAD ahahah
- from 2000 up to now I worked on my free book on Signals and TLC
  - I have started to translate it as background matherial for this class
- This is my third year teaching this course

# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Oh well, I know nothing about you!

Probably it's time to have a survey

- how many of you are foreign students?
- how many are enrolled in the third year?
- how many also attend the algorithm course?
- how many are unsure whether or not to choose this optional class?
- how comfortable are you with maths?
  - what about complex numbers and trigonometry?
- what are your expectations from this class?

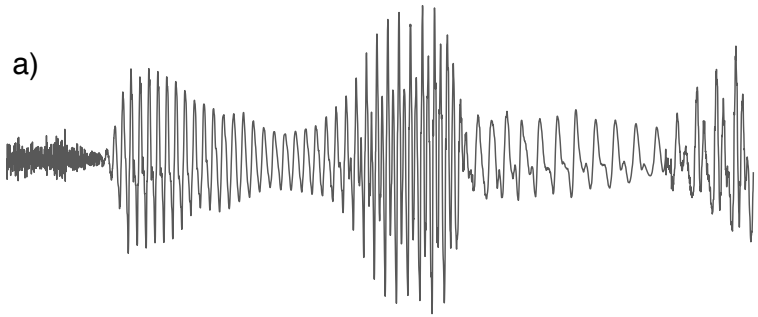
# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

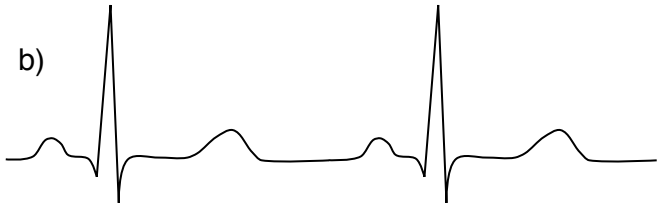
# Mathematical roots, and frequency analysis

## In the beginning it was maths

- Signals are nothing but functions $y = f(x)$, where
  - the independent variable is time, space, or other physical dimensions
  - the dependent variable is something that carries information




speech




electrocardiogram

- The waveforms above do not have an analytical expression
  - in signal theory they are called *stochastic processes*, unknown before they occur, but concrete after they had
  - even if devoid of analytical expression, they still have values
  - even with different waveforms, all signals of the same class (or process) share the same *frequency spectra*

# Mathematical roots, and frequency analysis

## Fourier transform or frequency content

- Very often things studied for one reason work well for another
  - J.B.P. Fourier developed its TRANSFORM as a solution for an heat conduction problem (1822)
  - put simply, the ***FT*** says that any waveform can be obtained as a weighted sum of sines and cosines, the amplitudes of which make up the frequency spectrum of the signal
- Signals analysis has been mainly studied in electronics, telecommunications, and sensing
  - its application to biology has developed over the past 25 years only
  - e.g. DNA is a *sequence* signal, meaning that both the independent (base number) and dependent (nucleobase id) variables are *discrete*
  - two completely different fields of knowledge have come into contact, and more is yet to come

# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Information content

## A measure of uncertainty

- As we said, a signal is such a thing only when it carries information
  - a steady waveform is always the same as itself, therefore its information content is limited to its *amplitude*, *frequency* and *phase*
  - a *road signal* we encounter along the way tell us something about what's to come
- The less likely the values of a signal or the elements of a sequence are, the greater their informational contribution
  - likelihood means probability
  - when the probability of the values of a signal depends on its past values, the information source is said *to have memory*
- The telecommunication aspects related to signal information are
  - transmission of a signal with minimal resource demand without losing information (called *source* coding)
  - transmission of a signal with added redundancy, so that errors can be detected or even better, corrected (called *channel* coding)

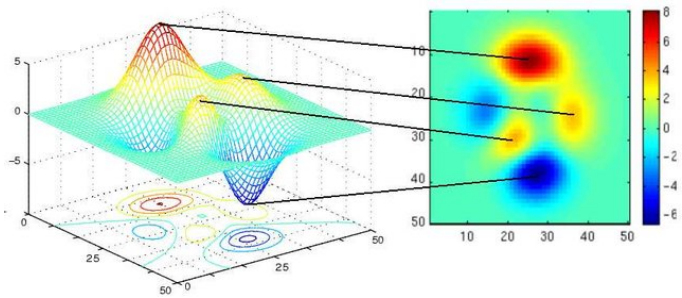
# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Audio, video, transmission and sampling

## Canonical fields of application for signals

- Audio informative signals are typically made up of speech and music, but also of sounds from nature and those generated by man-made machinery
  - let's listen and watch! (Audacity, Friture, VLC, Dfilter)
  - the most frequent processing is *filtering*, i.e. a change in freq. content
- Images are 2D signals (i.e. *surfaces*), with a couple of independent variables being spaces, and one (B&W) or three (RGB) independent intensity values



- Videos are sequences of images
- Both of the above can be generated or recorded (i.e. measured)
  - yeah, by now also synthetic images do exists!
- Obviously they can be transmitted by other types of signals (radio) or converted to a digital stream of bits (sampling and quantisation)

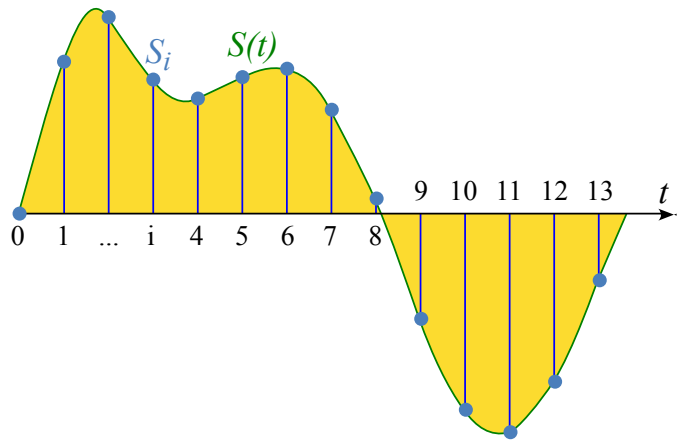
# Now we talk about...

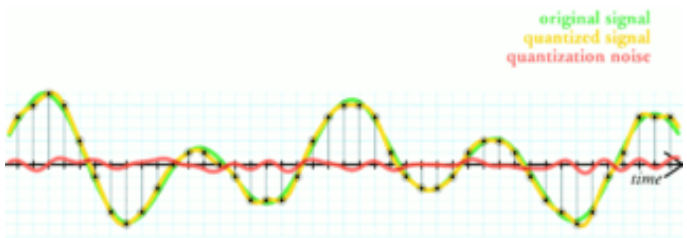
1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# The digital revolution

## From a continuous time signal to a discrete sequence in both time and values

- In the first half of the 1900s it became evident that a signal with limited frequency content can be exactly reconstructed from the mere knowledge of its regularly spaced samples
- its amplitudes must be *encoded* by a limited number of values and represented in binary mode
  - the fewer bits per sample used, the greater the *quantization noise*



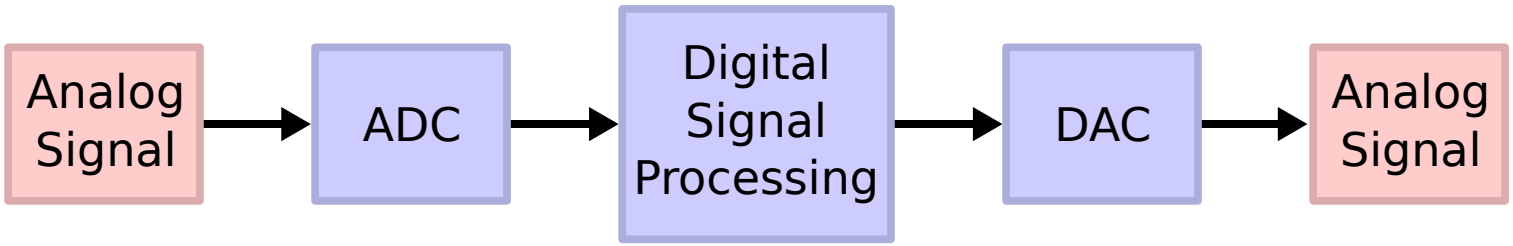




- in this way the signal can be written to CD, DVD, PC, pen drive, or even modulated and transmitted digitally
- it wasn't until 3/4 of the 1900s that technology became able to use the trick in real time

# The digital revolution

## Fourier also went digital, and Digital Signal Processing (DSP) was born

- Operations (e.g., filtering) defined in the analog domain are performed on digital bitstreams, and then converted back to the analog domain




- The Fourier transform that operates on a finite number of samples (or sequences) becomes a **DFT** (*Discrete Fourier Transform*) or (more efficiently) a **FFT** (*Fast FT*)
  - no chipset has a little man inside that calculates integrals, but a subunit that evaluates sum of products

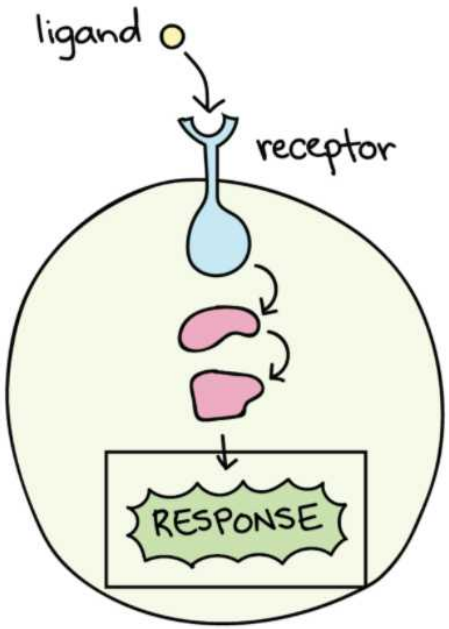
# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Biologist's concept of cellular signaling

## A matter of terminology

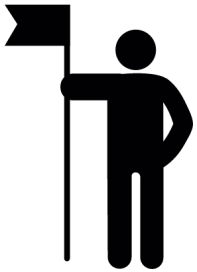
- The lexicon of the biologist has evolved independently of that of the communications engineer, and the same term has come to mean something different, albeit related
- The biologist calls cell signaling *the ability of a cell to receive, process, and transmit signals with its environment and with itself*
- It occurs thanks to the role of ligand molecules that interacts with the receptors present on the cell surface or its interior, thus initiating enzymatic activity, or opening or closing ion channels, or activating transporters and transcription factors, which in turn act by modifying the binding and localization properties of DNA, and modify some levels of gene expression
  - *as I understand it!*




# Signalist's concept of biochemical signaling

## From cell signalling to informational signals

- Taken literally, this way of telling the story seems to interpret biological signals in a way similar to a flag, or a single binary on/off switch
  - although descriptive, it does not at all represent the evolution of the values of a biological parameter along the time axis, or across the spatial dimension!

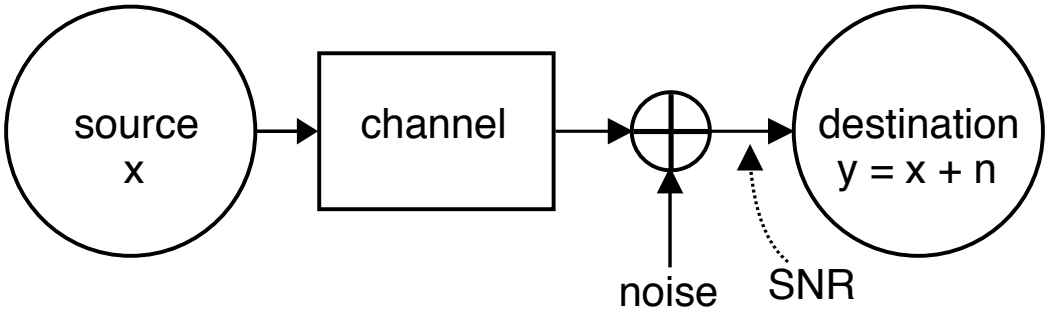


- More properly, cell signaling can be seen as an *information signal* that describes how
  - the concentration of the responsible ligand changes over time, thus making the related biological response more intense, i.e. expressed by more cells
- in this sense also the cellular response, i.e. the temporal variation of gene expression, can be thought of as an information signal, and
- the underlying chemical reaction represents a *communication channel*, in the telecommunications way of thinking

# The communication channel

## Information transmission from a telecommunications point of view

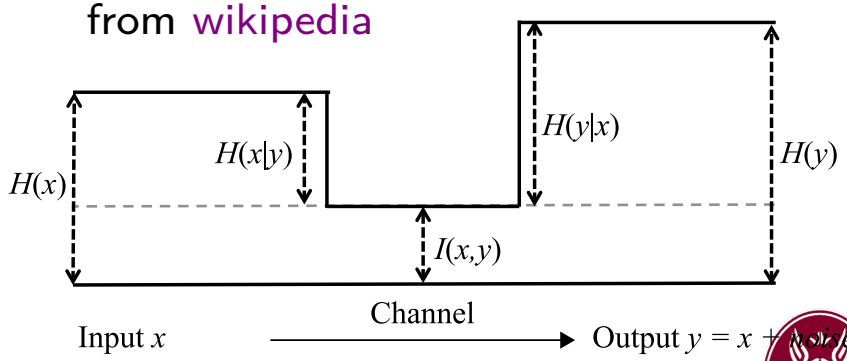
We can show it in a graphical representation as a conceptual flow




The presence of noise represents any type of disturbance, from *quantization noise* to *thermal disturbances* that add to the electrical signal, up to the *interference* produced by other communications, and to the *imperfect response* of the channel

- Signal to Noise power Ratio (SNR) is a measure of *transmission quality*, the higher the better
- the effect of noise is to cause *randomness* to the output, thus limiting the *mutual information* between the messages transmitted and received

from wikipedia

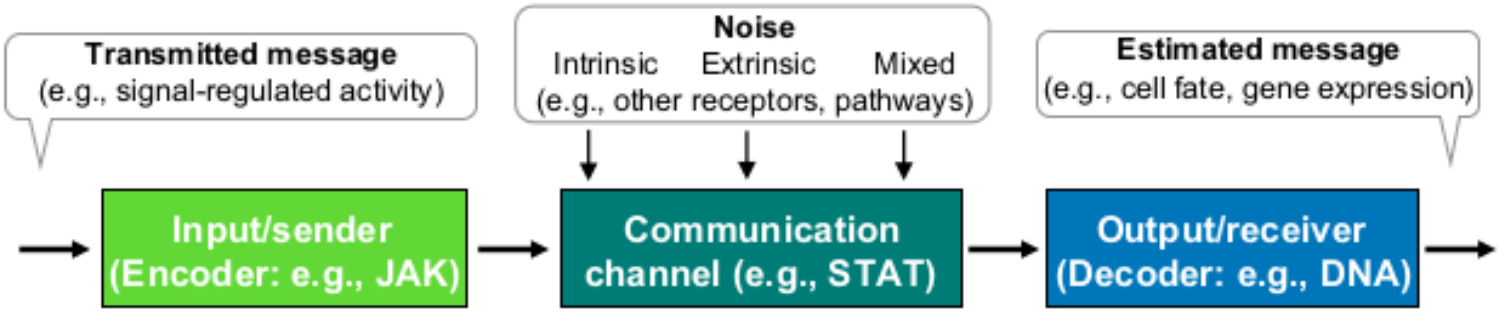


# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Loss of regularity (increased entropy) causes disease

There are some attemps to adapt *statistical communications theory* to biological signaling pathways, as for JAK-STAT




Just as an *increase in noise* in the communication channel

- decreases the *statistical dependence* between the sent and received messages
- consequently reducing the *mutual information* in between them
- making what is received have a *different meaning* from the original one

so

- a deviated behavior of a metabolic pathway can produce *undesirable results*

A collection of papers about this topic can be found here

# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# The cell cycle is periodic

## Finding the genes that drive mitosis

- 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
  - mRNA concentrations were used for the computation of a so-called *Fourier Score* as a measure of the periodicity of gene expression
  - here is an example of *lexical confusion*: after many thoughts, the *Fourier score* is actually the *amplitude of the first DFT coefficient* of the expression profile for the gene that transcribes to the mRNA
  - in this sense, it detects whether the expression profile makes *a complete change* during the time interval of a cell's division cycle

# The genome (and proteins) are sequences

## From letters to numbers

DNA is a chain made up of four letters (bases):

- A(denine), T(hymine), G(uanine) and C(ytosine)

Eukaryotic DNA contains genes

- divided into fragments called *exons*,
- separated by non-coding regions (*introns*),
  - which are eliminated by *splicing* during *transcription* from DNA to RNA

RNA *translation* makes *proteins*, long polymers of (20) *amino acid*

- students without a biology background can refer to this

Groups of three bases, called *codons*, do *code* for one specific a.a.

- the code is *redundant*, as a group of 3 elements with four values could code $4^3 = 64$ amino acid

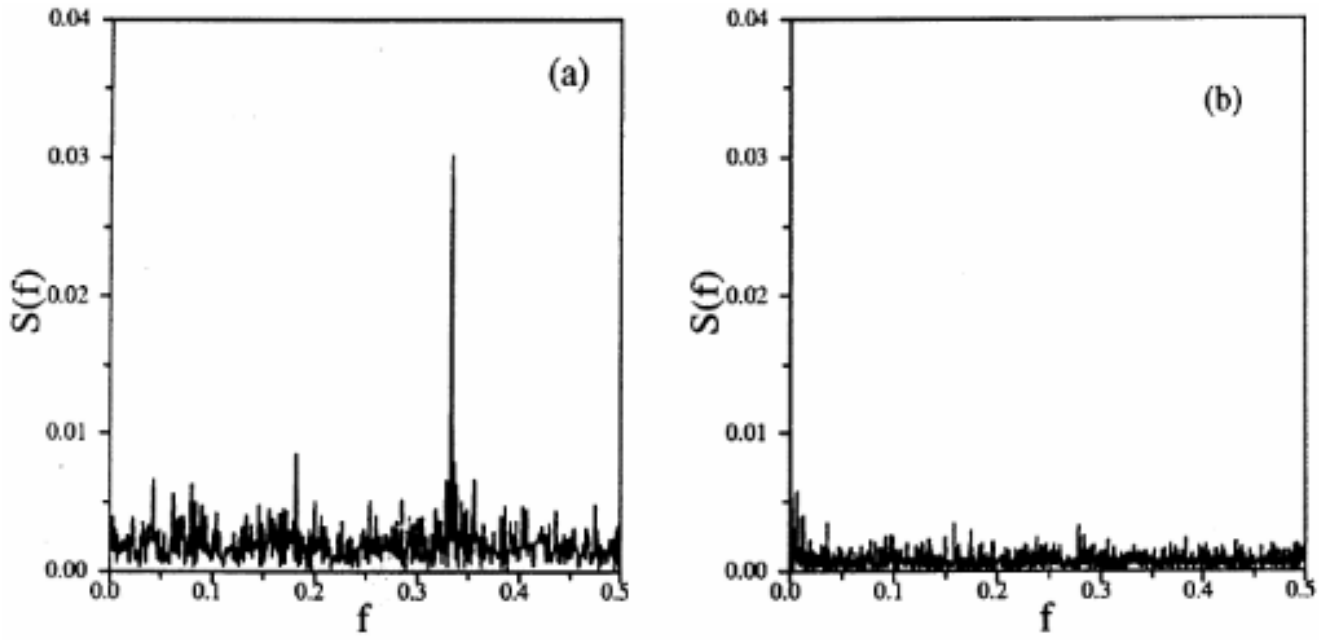
To apply DSP techniques, the DNA base chain must be *numerically encoded* as a *sequence* of numbers. Two approaches:

- use four different binary sequences (*indicators*), one for each base, indicating whether this is present (1) or not (0), or
- assign each of the four bases a different (real or complex) number, so that the nucleotide string becomes a single numerical sequence

# The DFT of the genome finds where the exons are

## Spectral analysis of a window from the numerical encoding of DNA

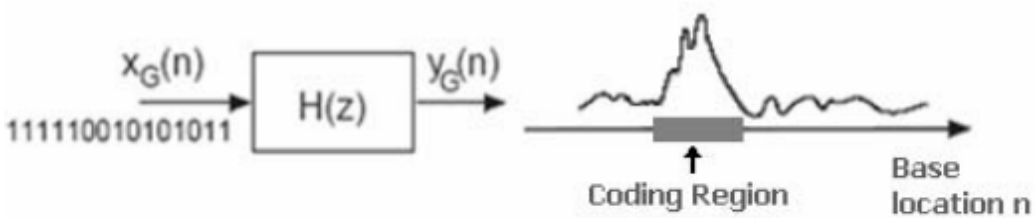
- The numerical encoding of DNA
  - is divided into segments of finite length, called *windows* in SP
  - for each window a *Discrete Fourier Transform* (DFT) is calculated, and
  - a frequency spectrum is associated to it
- If the DNA encoded window is *within an exon*
  - its spectrum shows *a sharp peak* at 2/3 of its bandwidth (link)

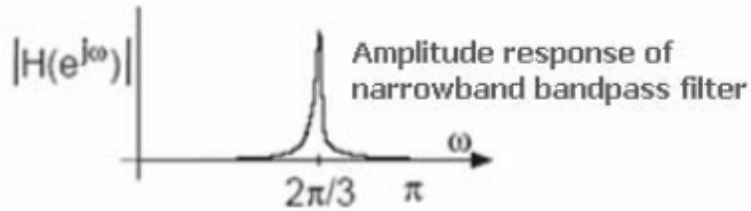


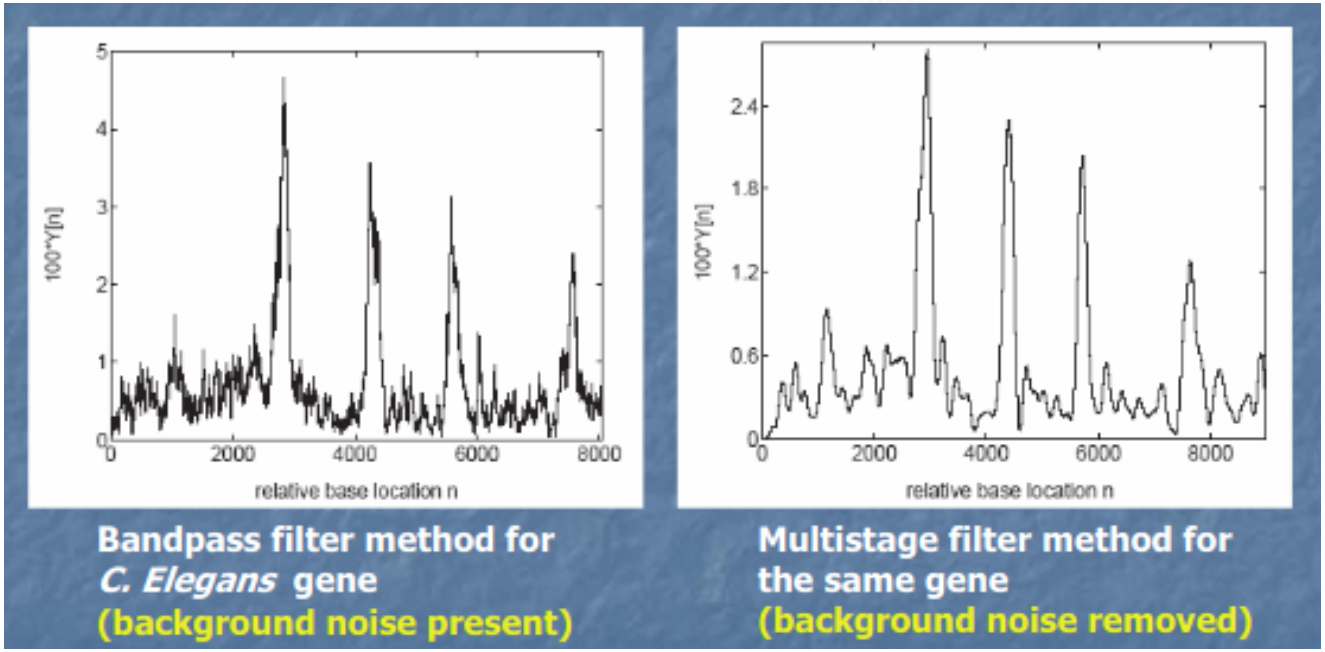

coding frame non coding frame

# Localization of DNA coding regions by filtering

- If the numerical sequence representing DNA is input to a bandpass filter
- then the location of its coding regions can be located by observing the output of the filter




# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Typical applications of SP in clinical routine

## Topics studied in clinical engineering courses

- The first applications of signal processing in medical-biological contexts emerged in the field of *diagnostics*, such as
  - Electrocardiography, electroencephalography, evoked potentials and electromyography - in which the electrical activity of the body is recorded and displayed
  - Doppler ultrasonography - a sort of radar for blood vessels
  - Tomographic reconstruction - used in CT scan and Positron emission tomography, it is based on an application of the Fourier transform, the Radon transform
- Some of the imaging techniques mentioned above can also be applied on a molecular scale

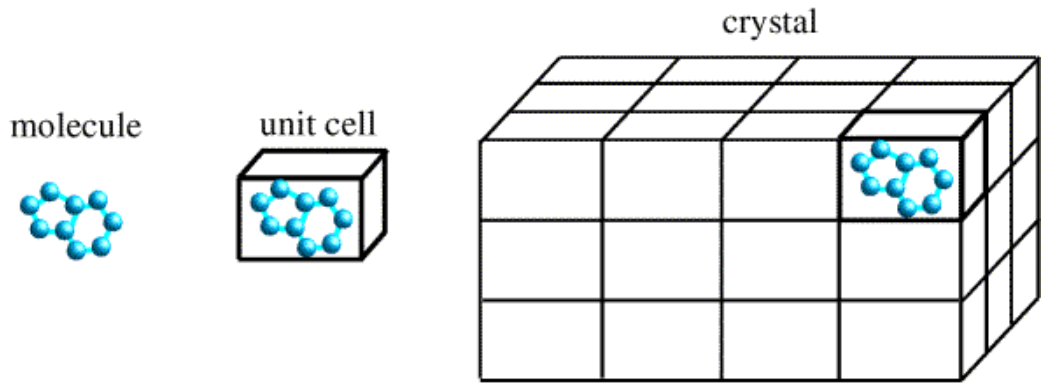
# Now we talk about...

1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# X-ray crystallography

## An inverse problem

It consists in obtaining the *diffraction pattern* of a monochromatic X-ray beam striking *a crystal* consisting of regularly spaced repetitions of the same protein molecule




- This topic is relevant for this course because
  - from the observed diffraction pattern we can directly derive the *coefficients* that define a *three-dimensional* Fourier series, and
  - these coefficients are those that could have been calculated directly starting from knowledge of the *charge density* of the valence electrons of the crystallized molecule
- In reality the electron density *is not known*, but on the contrary
  - it is obtained by calculating the Fourier series using the coefficients *estimated* from the diffraction pattern

# X-ray crystallography -2

## A physical and electromagnetic approch

The crytallography technique evolved in times before the affirmation of *signal processing*, and the relationship between the diffraction pattern and the Fourier transform of the electrons has a rather cryptic setting, which is reported below

- the incoming wave can be expressed as $Ae^{j\boldsymbol{k}_{in}\cdot\boldsymbol{r}}$ where $\boldsymbol{r}$ is a position in space, $\boldsymbol{k}_{in}$ is the wave vector, and $\boldsymbol{k}_{in}\cdot\boldsymbol{r}$ is a *dot product*
- the wave is scattered by the electron density $\rho(\boldsymbol{r})$ around the atoms, producing a spherical wave of amplitude (in direction $\boldsymbol{k}$) proportional (by $S$) to the # of scatters in a $d\boldsymbol{r}$ volume, i.e. $SAe^{j\boldsymbol{k}\cdot\boldsymbol{r}}\rho(\boldsymbol{r})\,d\boldsymbol{r}$
- a screen placed at $\boldsymbol{r}_{screen}$ receives the sum of all the scattered waves, each with wave vector $\boldsymbol{k}_{out}$ and phase change $e^{j\mathbf{k}_{\text{out}}\cdot(\mathbf{r}_{\text{screen}}-\mathbf{r})}$, so

$$diffraction(\boldsymbol{r}_{sceen}) = ASe^{j\mathbf{k}_{\text{out}}\cdot\mathbf{r}_{\text{screen}}}\int\rho(\mathbf{r})\mathrm{e}^{j(\mathbf{k}_{\text{in}}-\mathbf{k}_{\text{out}})\cdot\mathbf{r}}d\boldsymbol{r}$$

which, by putting $\boldsymbol{q} = \mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}$, takes the aspect of the 3D Fourier transform of the electron density $\rho(\boldsymbol{r})$, i.e.

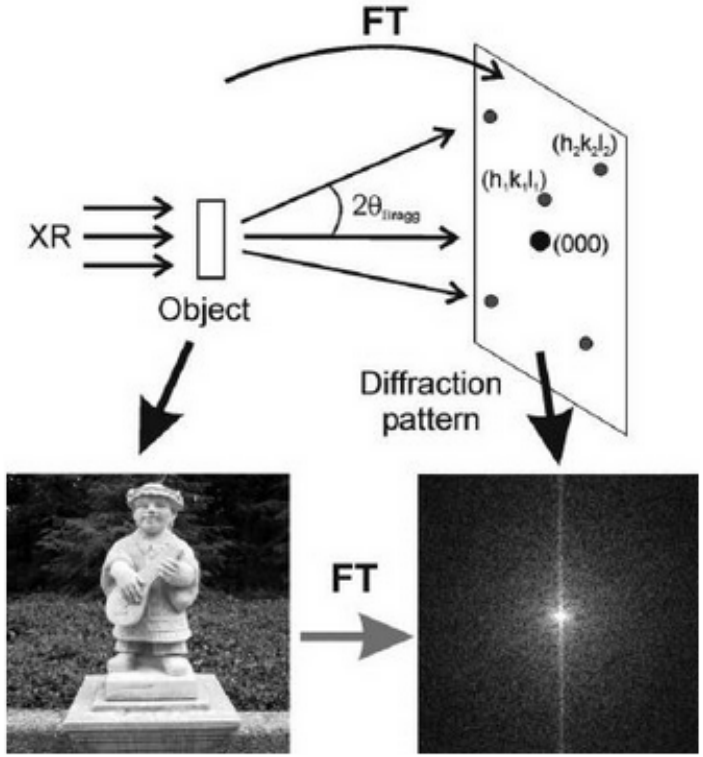
$$\int\rho(\mathbf{r})\mathrm{e}^{j(\mathbf{k}_{\text{in}}-\mathbf{k}_{\text{out}})\cdot\mathbf{r}}d\boldsymbol{r} = \int\rho(\mathbf{r})\mathrm{e}^{-j\mathbf{q}\cdot\mathbf{r}}d\boldsymbol{r} = F(\mathbf{q})$$

# X-ray crystallography - 3

## The signal processing approach and the phase problem

Fortunately, there is a way of looking at things that approaches 3D Fourier analysis as an extension of the two-dimensional case, very well studied for images, and which we will adopt in the course

- Incidentally, this will give us a good opportunity to discuss the basics of image processing too!

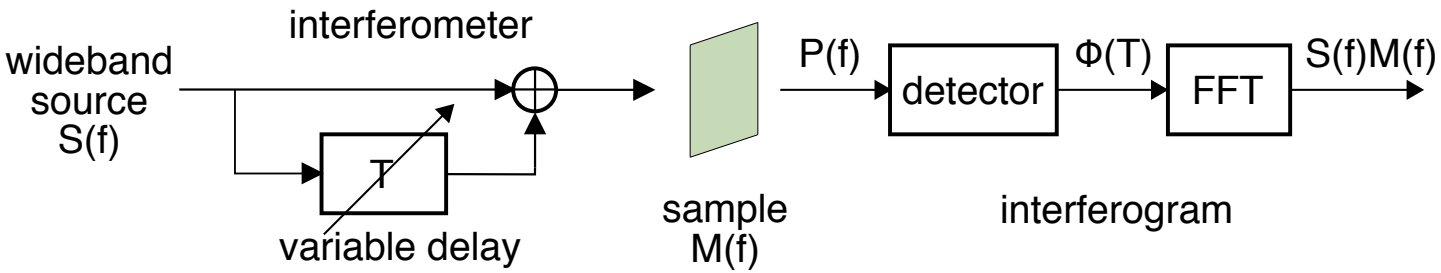



But.. this is only a part of the story...

- Actually, the measured intensity $I(q)$ of the diffraction pattern will be the square of the Fourier transform amplitude, i.e. proportional to $I(q) = |F(q)|^2$, thus losing the phase information
- The electron density $\rho(\boldsymbol{r})$ has to be retrieved by Fourier anti-transform of $F(q) = \sqrt{I(q)}$, i.e. $\rho(\boldsymbol{r}) = F^{-1}\{F(q)\}$, and the phase information put back by iterative methods

# Fourier transform infrared spectroscopy (FTIR)

- Spectroscopy is a means of determining the atomic composition of a material
  - based on the $\lambda$ wavelengths *of the light* it absorbs, i.e. doesn't let through
- This happens because
  - energy is converted to vibrations at resonant frequencies
  - which are associated with the normal modes of vibration and
  - depends on both the nature of the bonds and the mass of the atoms that are involved. Tables: [1] [2]
- It can be done by using all the $\lambda$ at the same time
  - by varying their mix as a function of time by means of an *interferometer*, which acts as a *comb filter*, and then
  - apply a Fourier transform




- More details later in the course

# Now we talk about...

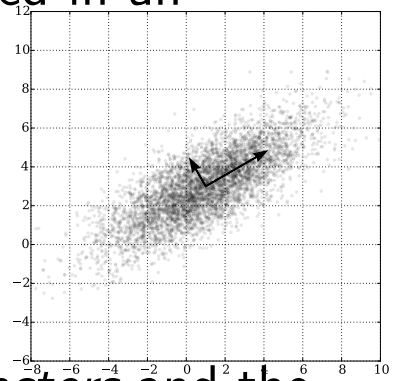
1. **Who we are**
   - Who am I
   - Who (I suppose) are you
2. **What signals are (up to yesterday)**
   - Mathematical roots, and frequency analysis
   - Information content
   - Audio, video, transmission and sampling
   - The digital revolution
3. **What is the role of Signal Processing in Bioinformatics**
   - Biologist's concept of cellular signaling, and its adoption
   - Loss of regularity (increased entropy) causes disease
   - The genome is a sequence, the cell cycle is periodic
   - Typical applications of signal processing in clinical routine
   - Crystallography and spectroscopy
   - Principal components, graphs and AI

# Principal Components Analysis and SVD

## But we won't deal with that

- The expression levels of $N$ genes in $M$ different contexts are recorded using microarray technologies
  - this can be thought of as *a cloud* of $M$ points scattered in an $N-$dimensional space

  We want to find a linear transformation, i.e. a coordinate change, which makes the new coordinates *not correlated*
- with each other, so that the new covariance matrix *is diagonal*
  - the whole thing is based on the search for the *eigenvectors* and the *eigenvalues* of the microarray covariance matrix
  - in the new space the coordinates are no longer individually related to the original gene set, so they are called *eigengenes* (not the song)
- Eigengenes can be *sorted* in a decreasing order of variance, and only those with the greatest variance are retained, thus obtaining a *reduction* in dimensionality
  - this is the *Principal Components Analysis* (PCA), based on the covariance matrix of the microarray, while the *Singular Values Decomposition* (SVD) works directly with the microarray matrix



# Signals on graphs

## and not even this other one

- The waveform of a time-dependent signal creates a dependency between values that follow one after the other
- An image creates dependencies between values placed in a grid
- The adjacency matrix of a graph creates a proximity relationship between values without the need for their linear or planar arrangement
  - we are faced with non-topological spaces
  - in this sense even a set of values connected by interaction relationships can be considered as a signal, and
  - concepts such as spectral analysis and filtering can be defined for it

# Neural networks

Often problems that

- cannot be addressed using linear techniques and models
- find solutions thanks to non-linear computational models based on statistical learning phases

But this course is already so full of things that we won't touch on this topic at all