Genomic Signal Analysis summary

P. Alexoudi, Prof. Ioannis Pitas

Aristotle University of Thessaloniki

pitas@csd.auth.gr

www.aiia.csd.auth.gr

Version 1.2

Artificial Intelligence &
Information Analysis Lab

Genomic Signal Processing (GSP)

- **Introduction to Genomic Signal Processing (GSP)**
- Introduction to Digital Signal Processing (DSP)
- Numerical representation of genomic sequences
- DNA string analysis
- RNA string analysis
- Protein string analysis
- 3D Protein Folding

Introduction to Genomic Signal Analysis

- ***Bioinformatics*** is a scientific field concerned with the use of computer science for the understanding of biological data (genome).

- Collection, organization and analysis of DNA and protein sequences.

Introduction to Genomic Signal Analysis

- The genomic information can be found as discrete sequences, whereas most signals in the environment appear to be continuous.

- ***Digital Signal Processing*** (***DSP***) uses digital processing, like computers and other processors, and mathematics to utilize the information signal and improve it.

- DNA and proteins can be represented as numerical sequences, therefore can be processed by DSP tools.

Introduction to Genomic Signal Analysis

- ***Genomic Signal Processing*** (***GSP***) is based on various disciplines and examines the processing of genomic signals.

- In particular, deals with the extraction of information from the gene, the analysis, process and use of the genomic signals produced in order to obtain valuable biological information.

- Has its roots is signal and systems theory.

Introduction to Genomic Signal Analysis

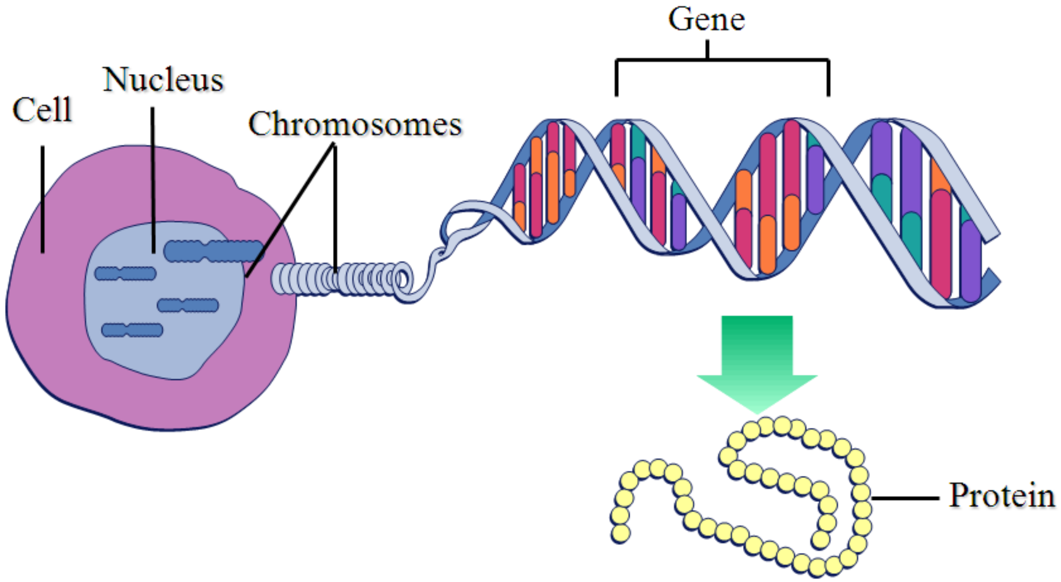
- All living organisms are consisted of cells.

- ***Genome*** can be characterized as the organism’s entire set of genetic information in a cell.

- The genome encompasses the instructions that are necessary to inherit in order to generate and preserve life and also reproduce.

Introduction to Genomic Signal Analysis

- In Eukaryotic organisms DNA is organized in ***chromosomes*** (e.g., human have 23 pairs of chromosomes).
- ***Gene*** is the basic physical and functional unit of heredity.
- Genes are segments of DNA. A particular class of genes are used to create molecules, characterized as proteins.



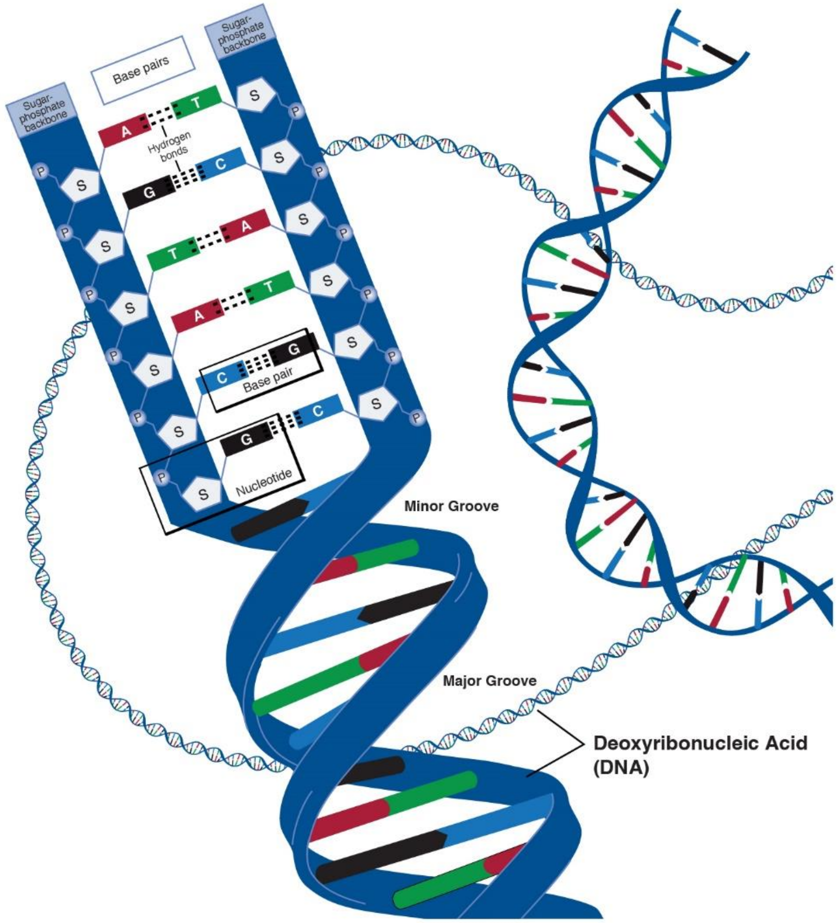
Reference: [GENE]

Introduction to Genomic Signal Analysis

- ***DNA*** (***Deoxyribonucleic acid***) is a molecule that carries the genetic instructions for all living organisms, even viruses.

- DNA is a double helix.

- Each strand of the helix is comprised of a sugar (deoxyribose) and phosphate.

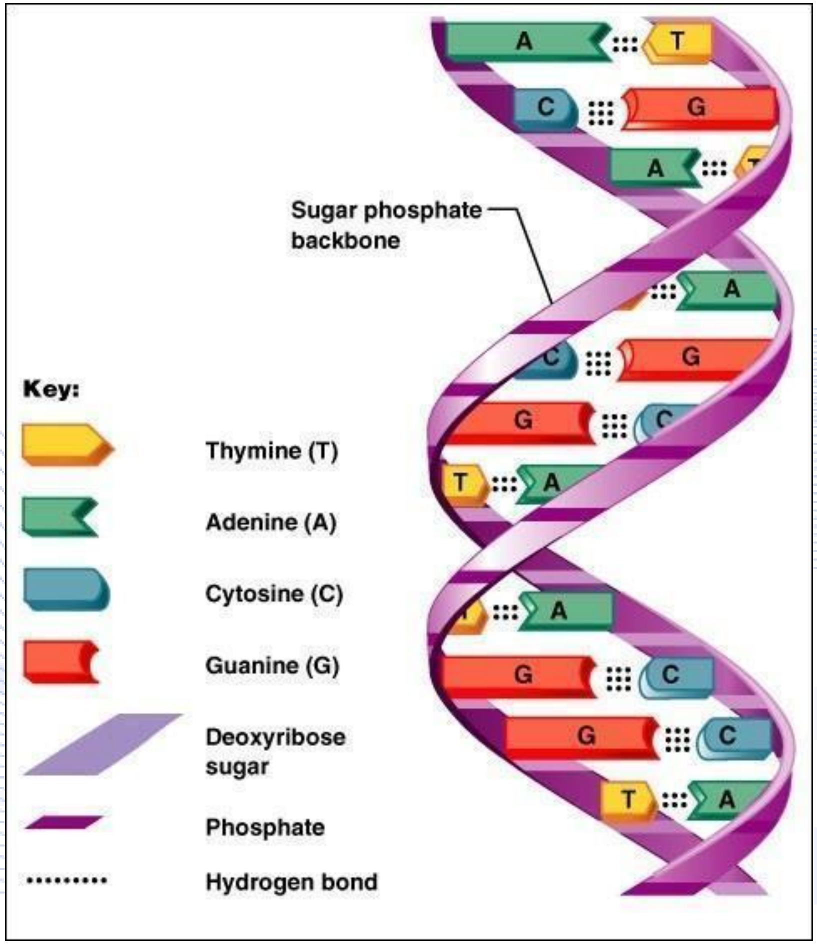


Reference: [DNA]

Introduction to Genomic Signal Analysis

- DNA has four bases, adenine (A), guanine (G), cytosine (C), and thymine (T).

- Adenine always pairs with Thymine and Cytosine with Guanine bonding the two strands.

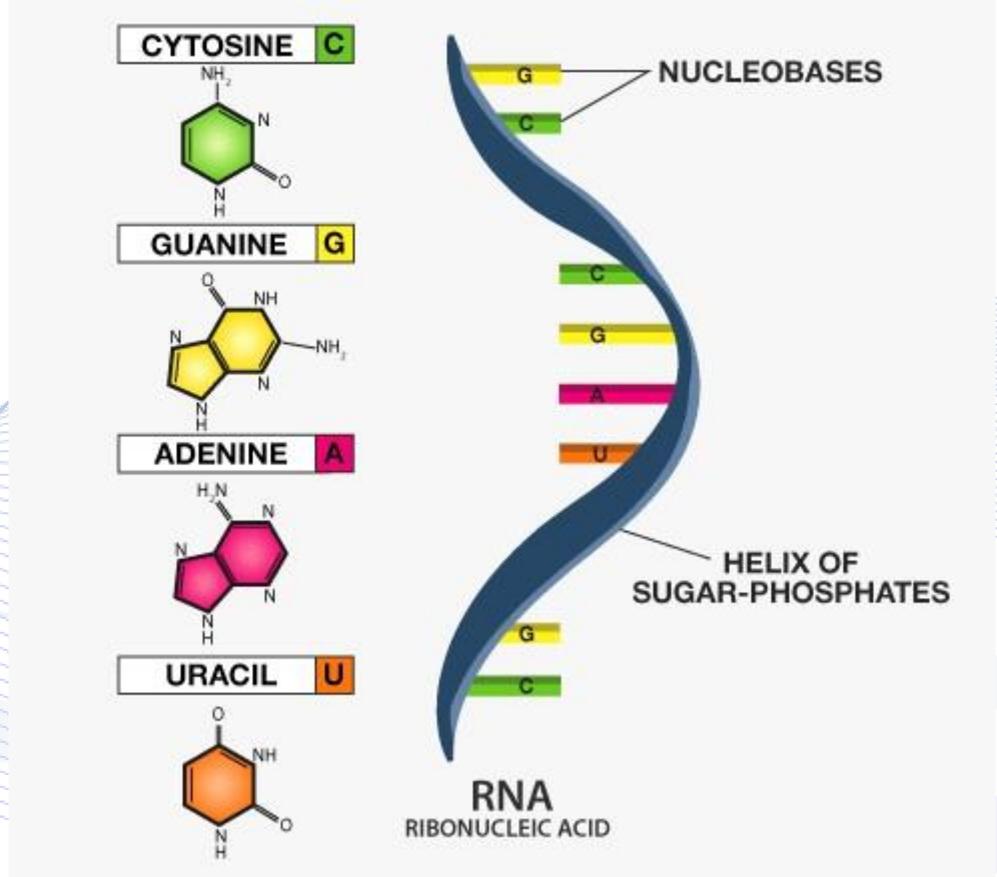


Reference: [ABD2017]

Introduction to Genomic Signal Analysis

- ***RNA*** (***Ribonucleic Acid***) is a molecule similar to DNA, but has a single strand.
- The sugar in the backbone of the strand is called ribose.

- In every sugar a base is attached, adenine (A), uracil (U), cytosine (C), or guanine (G).



Reference: [RNA]

Introduction to Genomic Signal Analysis

- RNA is responsible for coding, decoding, regulation and expression of genes.

- There exist three different types of RNA:
  - Messenger RNA (mRNA),
  - Ribosomal RNA (rRNA),
  - Transfer RNA (tRNA).

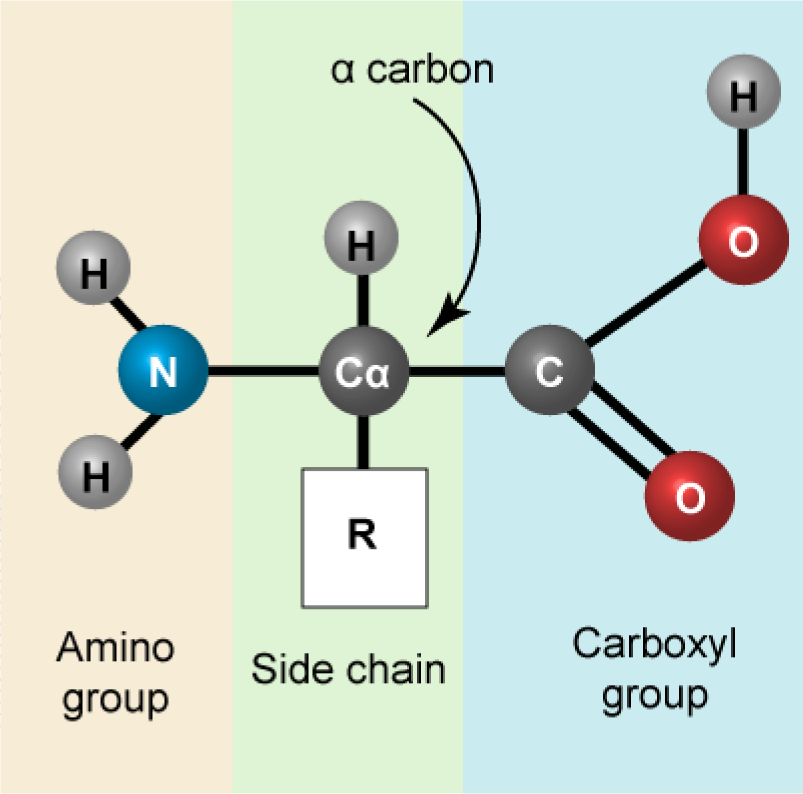
Introduction to Genomic Signal Analysis

- ***Proteins*** are large molecules that are comprised of hundreds or thousands of smaller units called amino acids.

- They are the building blocks of the cells and are necessary for the structure, function, and regulation of the body’s tissues and organs.

Introduction to Genomic Signal Analysis

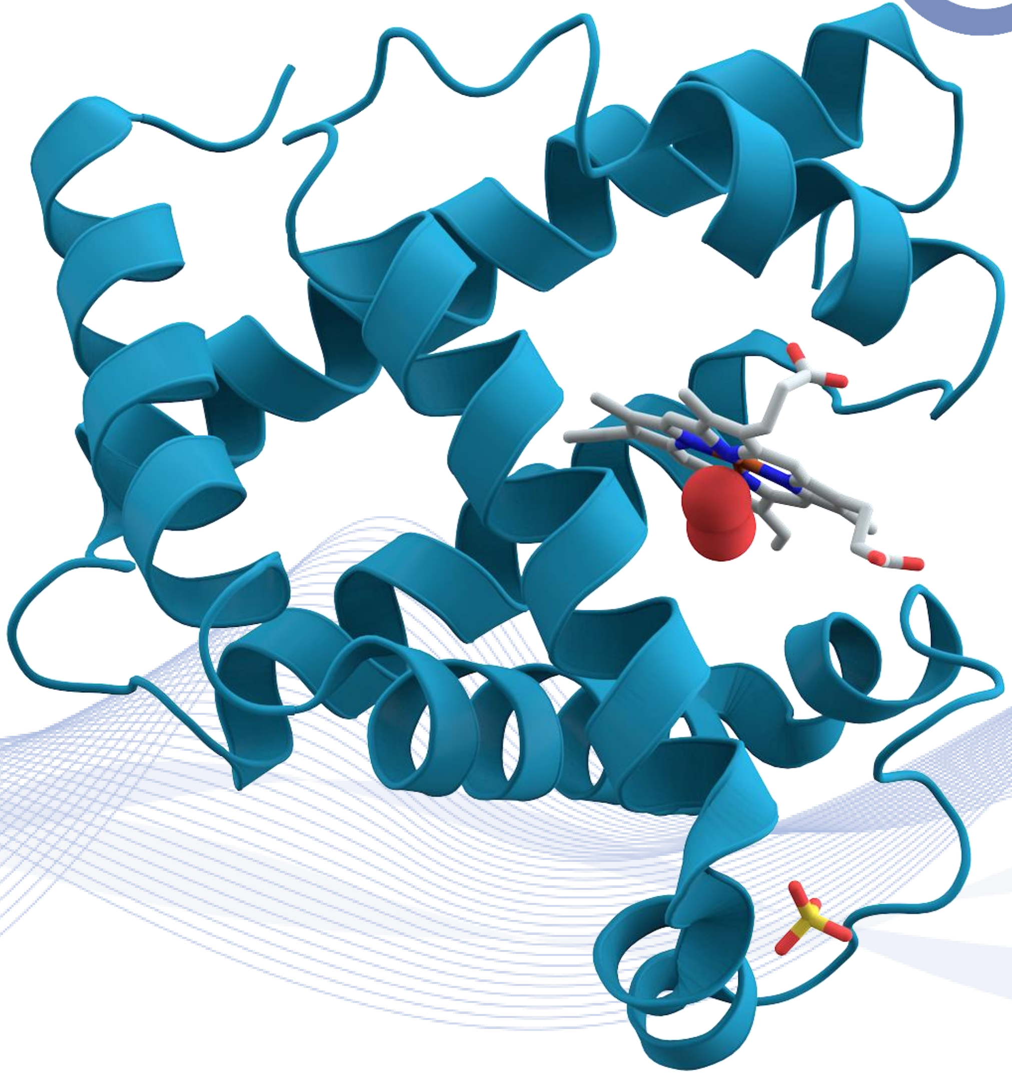
- ***Amino acids*** are monomers that create proteins.
- Consists of a central carbon atom ((α) carbon), attached to an amino group (NH2), a carboxyl group (COOH), a hydrogen atom and a variable side chain.
- Amino acids are divided into 20 different types and can be merged to form a protein.



Reference: [AMIN]

Introduction to Genomic Signal Analysis

- The sequence of amino acids give proteins the ability to fold into 3-Dimensional structures and perform their functionalities.



Reference: [PROTEIN]

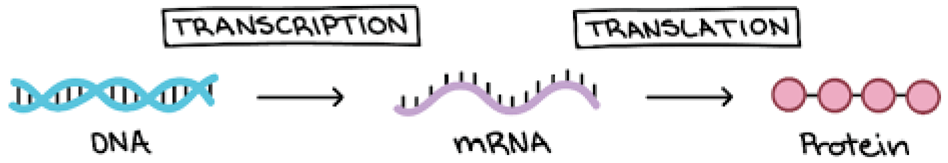
Introduction to Genomic Signal Analysis

- Functions of protein:
  - Transport and storage (e.g., Ferritin)
  - Messenger, regulate biological processes (e.g., Growth hormone)
  - Antibody, protect from foreign particles (e.g., Immunoglobulin G (IgG))
  - Structural component (e.g., Actin)
  - Enzyme, catalyze chemical reactions (e.g., Phenylalanine hydroxylase).

Introduction to Genomic Signal Analysis

- The ***central dogma of molecular biology*** is based on two fundamental procedures ***transcription*** and ***translation***.

- In transcription the genetic information is copied into messenger RNA (mRNA), while in translation the mRNA transcripts are responsible for the production of proteins.



Reference: [KHAN]

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- **Introduction to Digital Signal Processing (DSP)**
- Numerical representation of genomic sequences
- DNA string analysis
- RNA string analysis
- Protein string analysis
- 3D Protein Folding

DSP Algorithms for Genomic Sequences

- ***Discrete Fourier Transform*** (***DFT***) essentially transforms a discrete, finite function into another function:

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-i\frac{2\pi}{N}nk},$$

where $0 \le n,\ k \le N-1$.

- Magnitude and phase functions from the frequency spectrum of $x(n)$ can be used to depict DFT.
- Through DFT we can find periodicity in our data and in addition their relative intensities.

DSP Algorithms for Genomic Sequences

- The ***IIR digital filter*** is described by a finite difference equation of the following form:

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{k=0}^{M} b_k x(n-k),$$

where $x(n)$ is the input and $y(n)$output numerical sequence, a_k and b_k are numerical coefficients, n is the sample index, and k is an integer delay.

DSP Algorithms for Genomic Sequences

- The ***FIR digital filter***, on the contrary, is described by a convolution operation:

$$y(n) = \sum_{m=0}^{N-1} h(m)x(n-m),$$

where $h(m)$ is the impulse response of the filter.

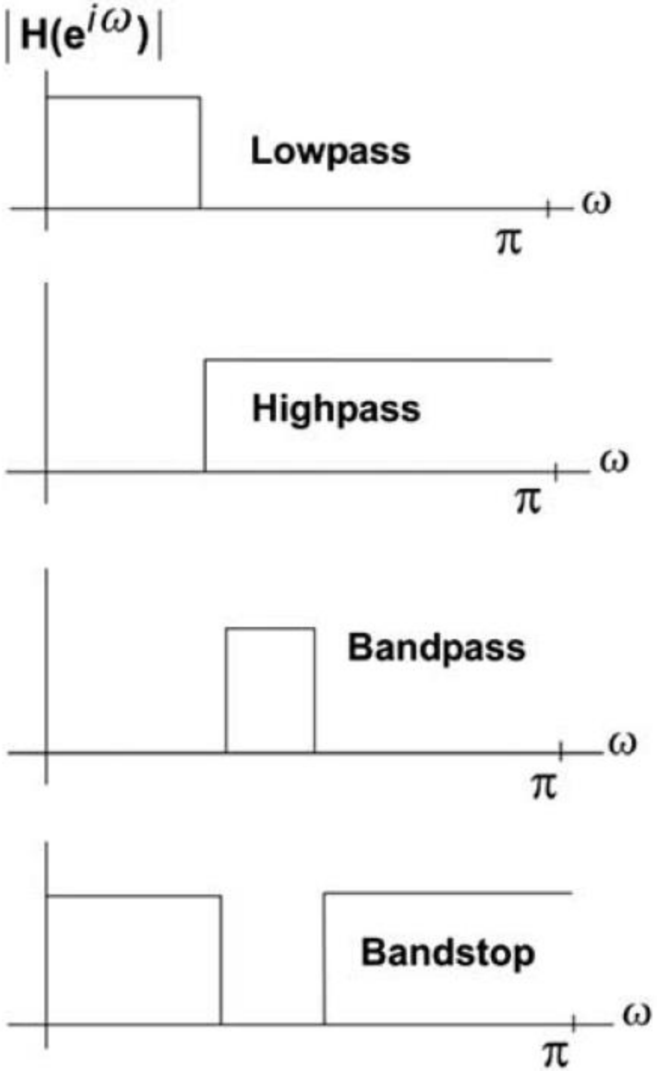
- A bilateral Z transform operator can be defined as:

$$Z\{x(n)\} = \sum_{n=-\infty}^{\infty} x(n)z^{-n},$$

where z is a complex variable.

DSP Algorithms for Genomic Sequences

- There are four basic prototype filter frequency responses for the magnitude based on the frequency band that is transmitted:
  - Lowpass.
  - Highpass.
  - Bandpass.
  - Bandstop.
- A multiband filter can be obtained by combining the above responses.



Reference: [LOR2009]

DSP Algorithms for Genomic Sequences

- ***Parametric models*** for spectral analysis have more advantages compared to other methods, like DFT.

- The PSD is determined by the parameters of the model and the variance of the input process.

DSP Algorithms for Genomic Sequences

- ***Entropy measures*** is a signal processing tool use in genomic sequences in order to measure randomness.
- The first definition by Shannon [SHA1948] for entropy is the following:

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i \,,$$

where p_i are the probabilities of the sequence $X = \{x_1, x_2, \dots, x_n\}$.

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- Introduction to Digital Signal Processing (DSP)
- **Numerical representation of genomic sequences**
- DNA string analysis
- RNA string analysis
- Protein string analysis
- 3D Protein Folding

Numerical Representation of Genomic Sequences

- The numerical representation of genomic sequences is important in order to utilize DSP techniques.

- One approach is through ***indicator sequences*** [VOS1992], where each base is represented with a binary sequence, with 1 suggesting the presence and 0 the absence of that base in a certain location.

Numerical Representation of Genomic Sequences

- The indicator sequences for the four bases of DNA are the following:

$$x_A[n] = 1000100010101\ \ldots$$
$$x_T[n] = 0011000100000\ \ldots$$
$$x_C[n] = 0100001001000\ \ldots$$
$$x_G[n] = 0000010000010\ \ldots,$$

where n is a finite state.

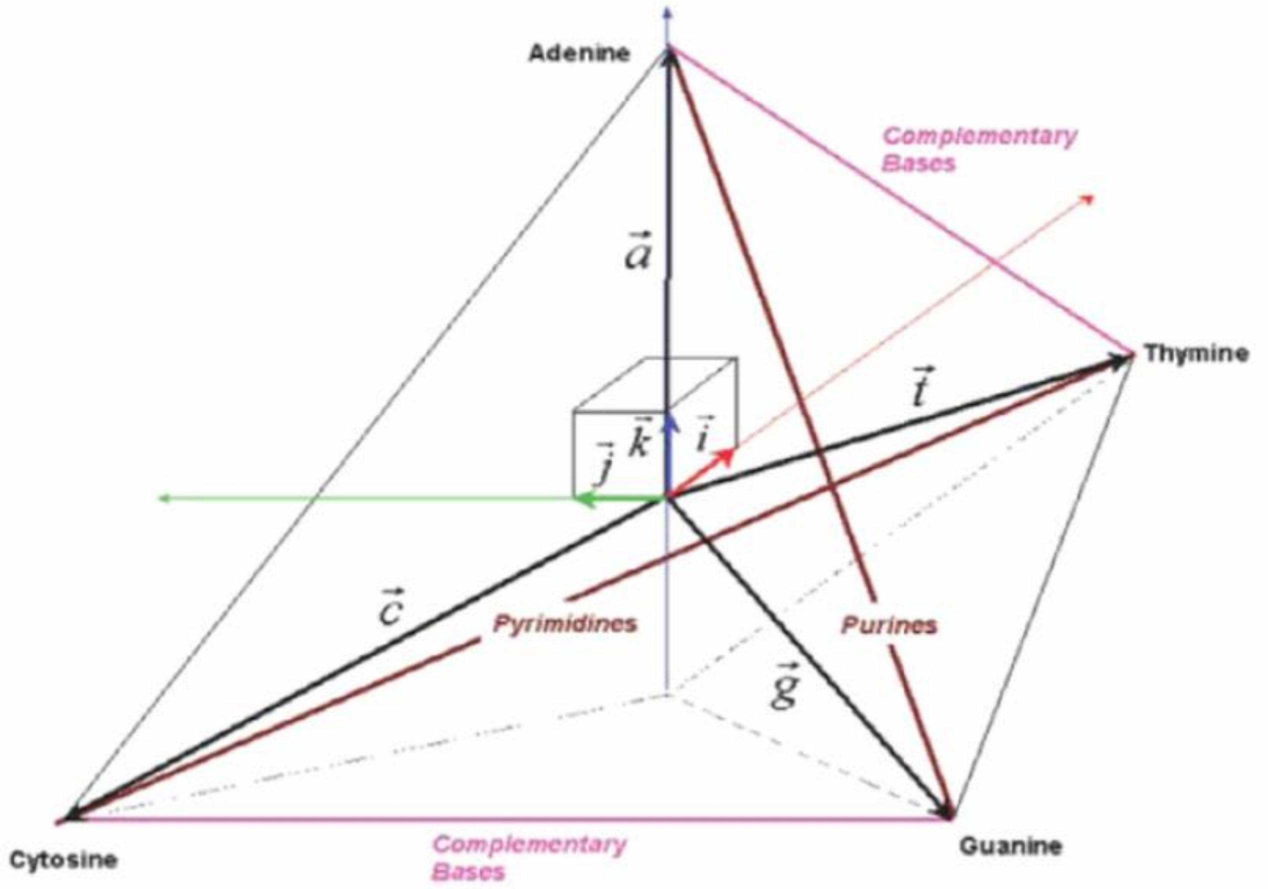
Geometric Representations

- A ***tetrahedral representation*** of nucleotides is proposed by [CRI2002].

- Each base A, T, G and C is assigned to the vertices of the tetrahedron with the length vectors being symmetric to each other and pointing to the corners.

- It is noted that the vertices of a regular tetrahedron are a subset of the vertices of a cube.

Geometric Representations

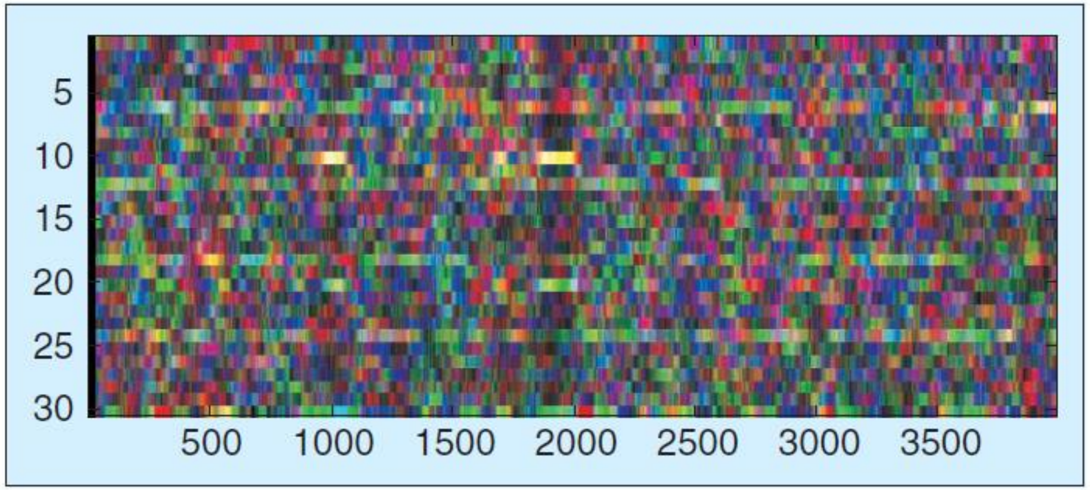


Tetrahedral representation of DNA bases.

Reference: [CRI2002]

Color spectrograms of DNA

- ***Spectrograms*** are the main application of tetrahedral representation useful visualization tools that can provide insights about the DNA sequences.
- DNA Spectrograms are defined by [ANA2001] utilizing indicator sequences as follows:
  - Obtain the three magnitudes of STFT that correspond to the primary colors Red, Green and Blue.
  - Superimpose the three STFT matrices.



Reference: [ANA2001]

Quaternion Representation

- The coordinates of the vertices of the cube are assigned to an integer $\{\pm 1\}$.
- The form of the base vectors is the following:

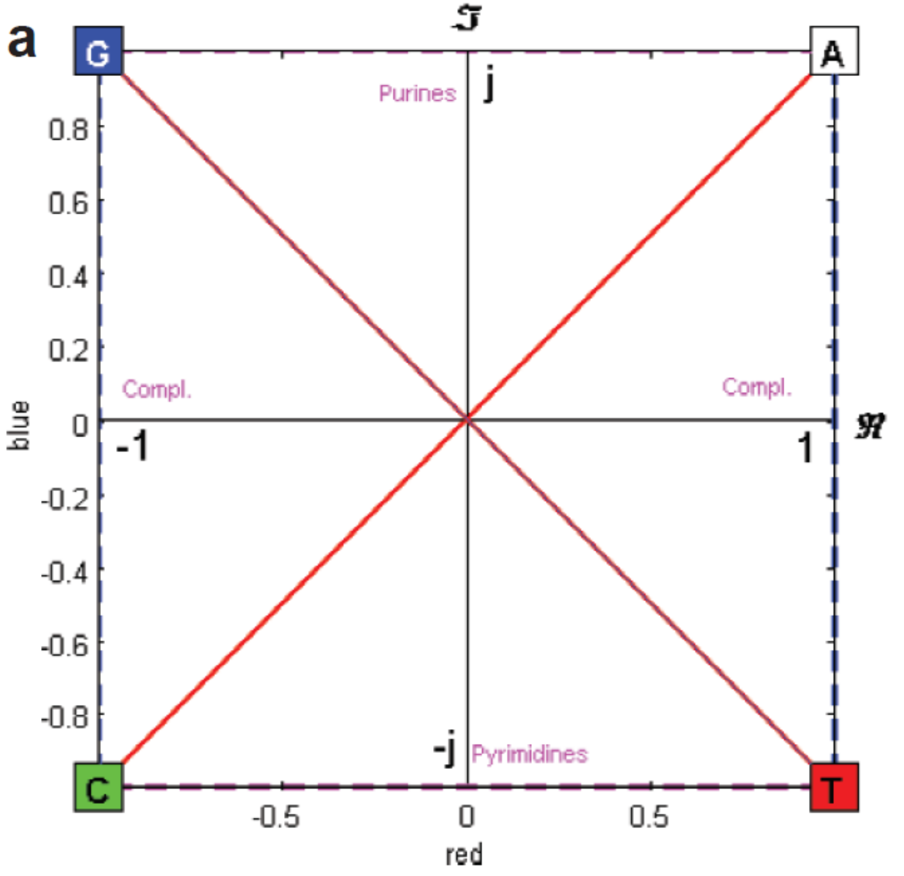
$$\boldsymbol{\alpha} = \mathbf{i} + \mathbf{j} + \mathbf{k},$$
$$\mathbf{c} = -\mathbf{i} + \mathbf{j} - \mathbf{k},$$
$$\mathbf{g} = -\mathbf{i} - \mathbf{j} + \mathbf{k},$$
$$\mathbf{t} = \mathbf{i} - \mathbf{j} - \mathbf{k}.$$

- The three primary colors of the RGB system are assigned to the axes of the system, therefore, each point corresponds to a certain hue.

Complex Representation

- The tetrahedral representation can be reduced to 2D, if the original tetrahedral is projected on a plane.
- There are various ways of choosing projection planes.
- For example, the red-blue (xz) plane as shown in figure form the following complex representation of the bases:

$$\boldsymbol{\alpha} = 1 + \mathbf{j}, \qquad \mathbf{c} = -1 - \mathbf{j},$$
$$\mathbf{g} = -1 + \mathbf{j}, \qquad \mathbf{t} = 1 - \mathbf{j}.$$



Reference: [CRI2002]

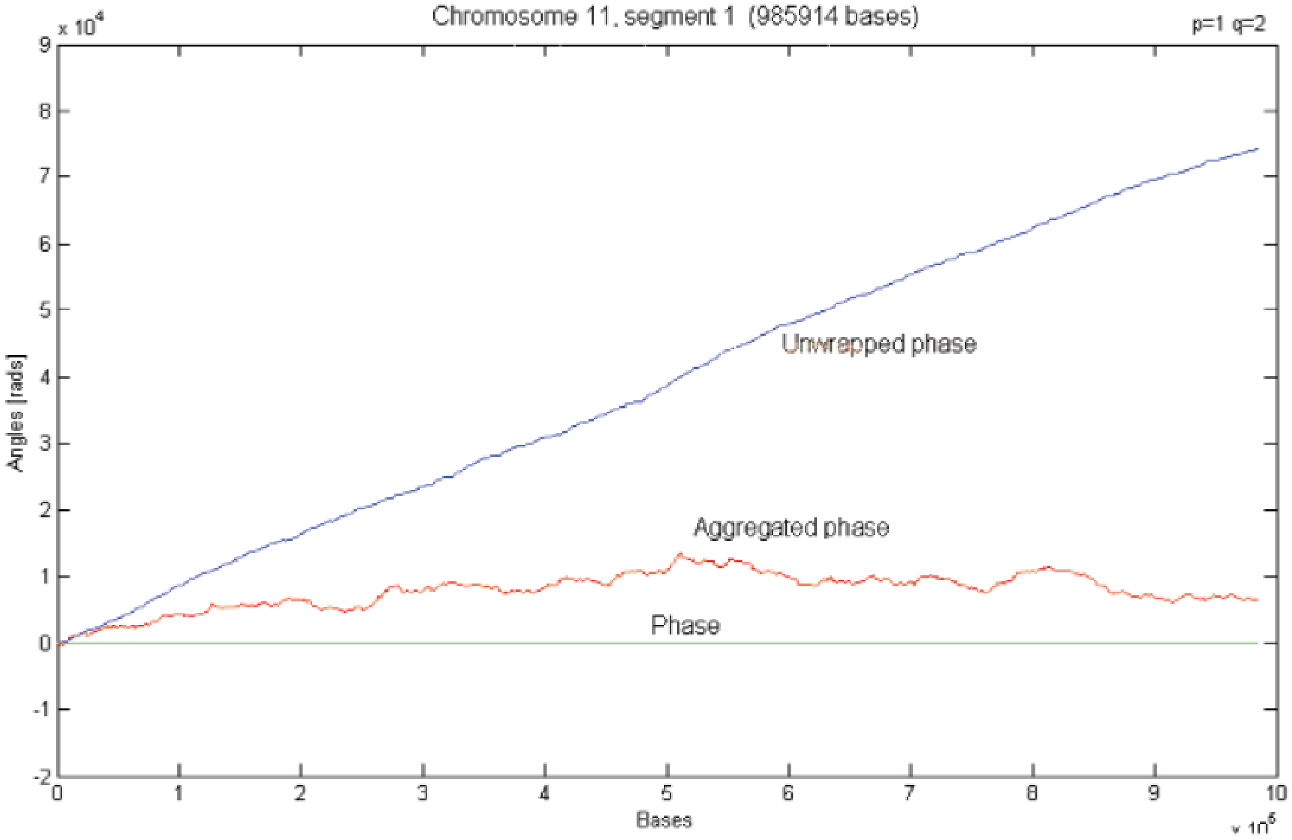
Geometric Representations

- ***Phase*** of a complex number is a periodic multi-valued magnitude.

- ***Cumulative or aggregated phase*** is the sum of the complex base representations starting from the first element of a sequence.
  - Its value is never zero, but drifts between negative and positive values suggesting the relative frequencies of bases.

Geometric Representations

- ***Unwrapped phase*** is the corrected phase of a sequence of complex numbers.
  - It can be calculated by adding or subtracting 2π to or from the phase of the new element.
  - Its value suggests the relative frequencies of the transitions between the bases.



Reference: [CRI2002]

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- Introduction to Digital Signal Processing (DSP)
- Numerical representation of genomic sequences
- **DNA string analysis**
- RNA string analysis
- Protein string analysis
- 3D Protein Folding

Long range correlations in DNA

- A ***long range correlation*** among base pairs of DNA sequences both for coding and non-coding regions exists.

- The autocorrelation for each indicator base was given in the section of parametric models.

- The $1/f$ behavior indicates a slowly decaying term in the autocorrelation sequence, which holds for the term long range correlation.

Long range correlations in DNA

- The existence of the $1/f$ behavior in the DNA spectrum might be explained by the duplication-mutation model.
- Li [LI1997] studies the correlations among the four bases, e.g.:

$$r_{AG}(k) = \sum_{n} x_A(n) x_G(n-k).$$

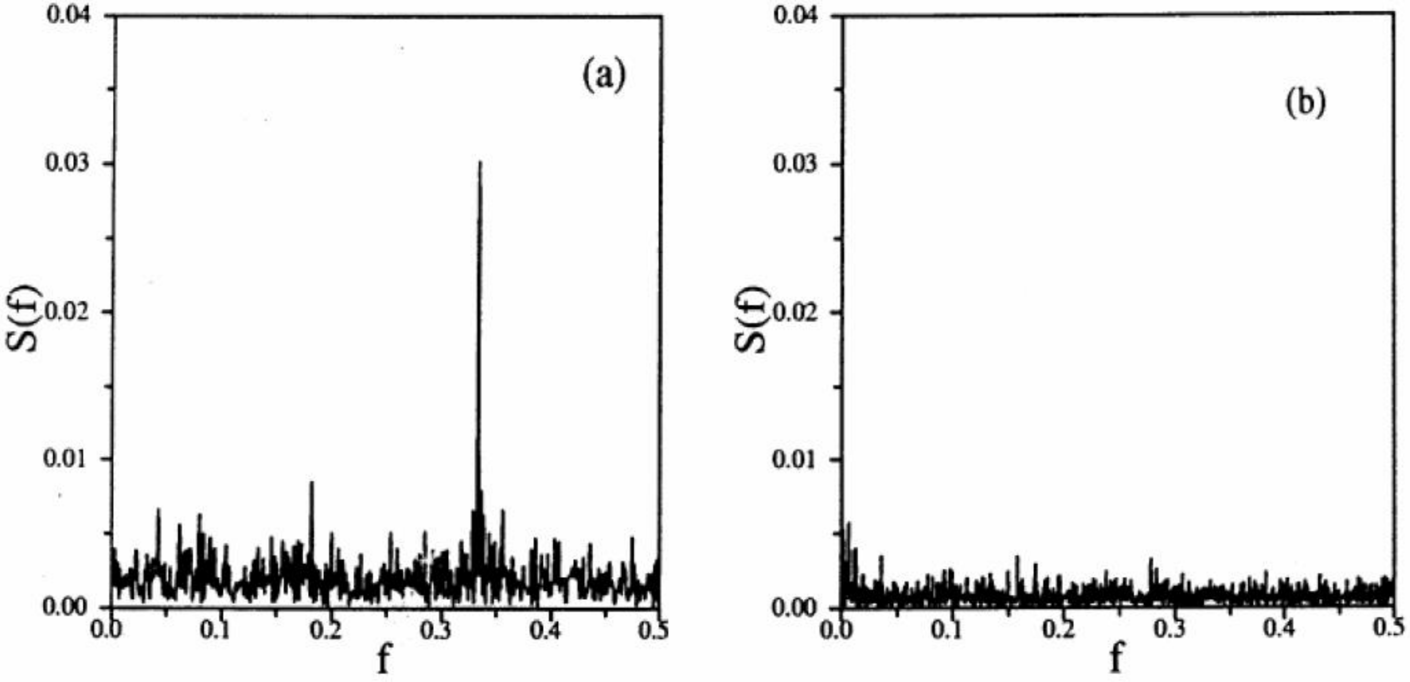
- This results in the following correlations:

$$r_{AG}(k) \approx r_{CT}(k), \;\; r_{AA}(k) \approx r_{TT}(k), \; r_{CC}(k) \approx r_{GG}(k).$$

Artificial Intelligence & Information Analysis Lab

Identification of protein coding DNA regions

- One of the first approaches to implement Fourier analysis in gene prediction using 3-base periodicity was proposed by [TIW1997]:



Reference: [TIW1997]

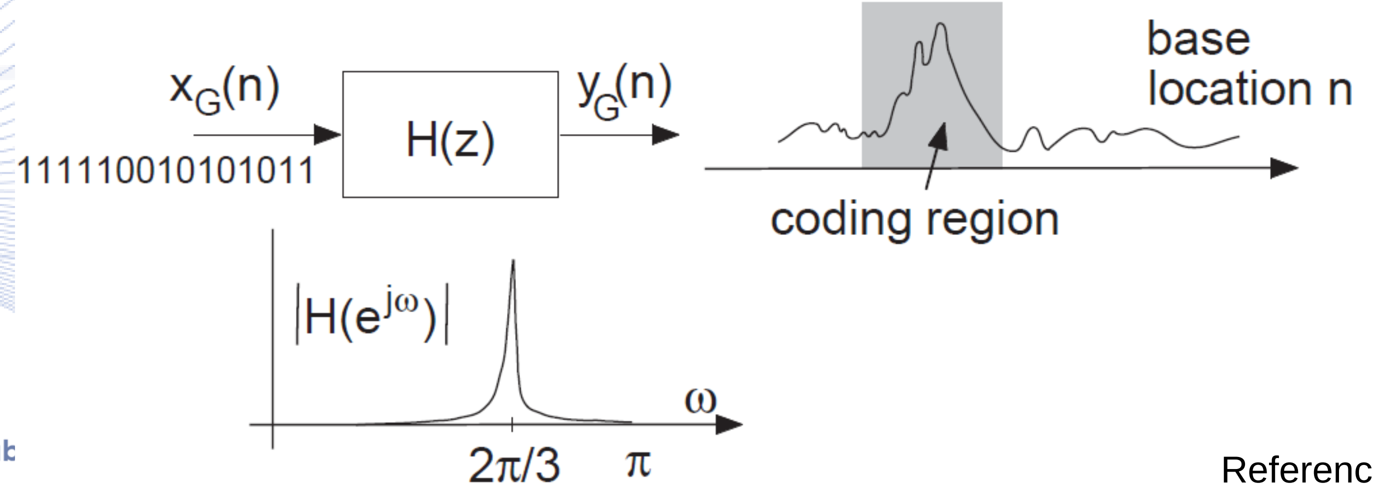
Power Spectral Density of (a) coding and (b) non-coding regions.

Identification of protein coding DNA regions

- The plot of the output $Y(n)$ can help identify coding regions:

$$Y[n] = |y_A(n)|^2 + |y_T(n)|^2 + |y_G(n)|^2 + |y_C(n)|^2.$$

- $y_A(n)$, $y_T(n)$, $y_G(n)$ and $y_C(n)$ are the output sequences, after applying the digital filter in the indicator sequences $x_A(n)$, $x_T(n)$, $x_G(n)$ and $x_C(n)$ and n is the base location.



Reference: [VAI2004]

Signal Extraction for DNA microarray

- Gene expression is usually calculated by the amount of mRNA expressed in genes.

- ***DNA microarray*** is a powerful tool that can be utilized to document gene expression in various levels as a function of time.

- Microarray technology includes complementary DNA (cDNA) and oligonucleotide.

Signal Extraction for DNA microarray

- Signal processing techniques, like normalization, clustering and denoising are used in the analysis of data obtained by microarray DNA.
- In the work by [ALT2000] microarrays are expressed using a matrix:

$$\mathbf{X} = [\mathbf{x}_{nm}], \qquad 0 \leq n \leq N-1, \qquad 0 \leq m \leq M-1,$$

with the columns being the expression levels of N genes and the rows the expression levels of a single gene at different times.

Alignment methods

- Another method for genomic analysis are ***alignment methods***, that determine the distances between different sequences so as to discover similarities in two or more sequences.

- The comparison of two sequences is called ***Pairwise Sequence Alignment*** (***PSA***), while when more than two are compared the process is called ***Multiple Sequence Alignment*** (***MSA***).

Phylogenetic analysis

- ***Phylogenetic analysis*** is the most important tool and application of PSA for DNA analysis.

- ***Phylogenetic trees*** give the ability to classify DNA sequences, organize information about biological diversity and provide information regarding evolution.

- Dichotomous trees can be used to depict the separation between organisms.
  - Branches that are close enough show similarity between organisms, whereas branches that are far away show large differences.

Machine Learning in Genomic Signal Analysis

- ***Classification*** has been vital in genomic signal analysis, for example in the analysis of microarray gene expression data.

- Classifiers provide information about the state of a cell from an expression vector and output a class label called phenotype to distinguish a healthy from a non-healthy, the type of cancer etc.

- Methods for classification of species belong into the two categories ***alignment-based*** and ***alignment-free***.

Classification

- ***Molecular Evolutionary Genetics Analysis*** (***MEGA***) is an alignment-based software proposed by [KUM2016].

- It provides a statistical analysis of gene sequences and focuses on evolutionary relationships and patterns of DNA and protein evolution.

- It is considered the state of the art software for alignment-based methods.

Classification

- An alignment-free method that combines supervised learning with DSP techniques is proposed by [RAN2019], called ***Machine Learning with Digital Signal Processing*** (***ML-DSP***).

- It is a tool for genomic DNA sequence classification that achieved above 97% accuracy in classifying genomes from different species.

Classification

- The feature vector of ML-DSP is based on the Pearson Correlation Coefficient (PCC) and is derived from the DFT magnitude spectra of the discrete numerical sequences.

- ML-DSP outperforms the state of the art alignment-based software MEGA7 regarding time processing.

Clustering

- ***Clustering*** has been particularly used for gene expression microarrays and as similarity computation methods.
- Clustering uses the expression vector to cluster data points into subsets.
- Clustering algorithms applied to genomic data are k-means, fuzzy c-means, self-organizing maps, hierarchical clustering, and model-based clustering.

Genomics

- Genomics include two main categories:
- ***Structural genomics***: Is comprised of genome sequencing, genome organization, EST sequencing, physical mapping, molecular cytogenetics, and linkage analysis.
- ***Functional genomics***: Is comprised of gene expression technologies, forward and reverse genetics, and comparative genomics.

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- Introduction to Digital Signal Processing (DSP)
- Numerical representation of genomic sequences
- DNA string analysis
- **RNA string analysis**
- Protein string analysis
- 3D Protein Folding

RNA string analysis

- ***Transcriptome*** is defined as the set of RNA molecules produced through transcription.

- Transcriptome constantly changes as it is affected from many factors, like the conditions of the environment.

- It is a powerful tool to define the basis of genome, using mRNA, non-coding RNA and small RNA.

RNA string analysis

- The initial goal is to define the transcriptional structure of genes, posttranscriptional modifications, splicing patterns and differential expression analysis.

- The procedure of transcription incudes:
  - mRNA, that is a ladder from gene to proteins.
  - Non-coding RNA (cRNA), that is in charge of control of gene expression.

- The above knowledge gives insights to the biological activity of gene.

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- Introduction to Digital Signal Processing (DSP)
- Numerical representation of genomic sequences
- DNA string analysis
- RNA string analysis
- **Protein string analysis**
- 3D Protein Folding

Protein string analysis

- A ***proteome*** is characterized as the entire set of proteins expressed in an organism.

- ***Proteomics*** is a scientific field that focuses on proteomes and can be divided into two categories:
  - Proteome analysis that focuses on comparisons to identify and localize proteins.
  - 3D protein structure.

Sequence alignment

- As mentioned earlier a powerful tool for analysis is sequence alignment.
- The comparison of primary amino acid sequences can provide information about their distance and correlation.
- The most popular algorithm used is BLAST and an improvement of it is Psi-BLAST (Position-Specific Iterated BLAST).

Phylogeny

- Again phylogenetic trees can be comprised by aligning their sequences.
- One common method to estimate protein distances is the minimum distance.
- Phylogenetic trees, through analysis of the changes in sequences, provide information about homologous proteins, the groups created due to similarity and their evolution, as well as their functional diversity.

Secondary structures

- Defining the secondary structure of proteins follows sequence analysis.
- The elements of secondary structure can be classified to alpha helices, beta strands and undefined structure sequences (coils).
- In order to predict the secondary structure there are three different types of methods.

Structure Alignment

- Structure alignment algorithms are comprised by the following steps:
  - Measure the center of mass for every structure.
  - Overlap the two structures (center of mass should be equal).
  - Measure the angles between residues using the center of mass as starting point.
  - Rotate one of the two structures using the median angle difference.
- An adequate step for more compound structure alignment is to create a distance score matrix.

Protein-Protein interactions

- Most of the proteins do not behave as individual units. They interact with other proteins to create more complex forms.

- The interactions between proteins are not just restricted in creating physical binding.

- Proteins can also being involved indirectly in large protein groups, in the regulation between them and share a substrate in a metabolic pathway.

Genomic Signal Processing (GSP)

- Introduction to Genomic Signal Processing (GSP)
- Introduction to Digital Signal Processing (DSP)
- Numerical representation of genomic sequences
- DNA string analysis
- RNA string analysis
- Protein string analysis
- **3D Protein Folding**

3D Protein Folding

- 3D protein structure is vital as it indicates the functionality of the protein.

- The protein folding problem, which is the identification of the shapes proteins fold into has been a challenge for nearly 50 years.

- The prediction of this structure gives insights into developing treatments for diseases and discover enzymes for many processes.

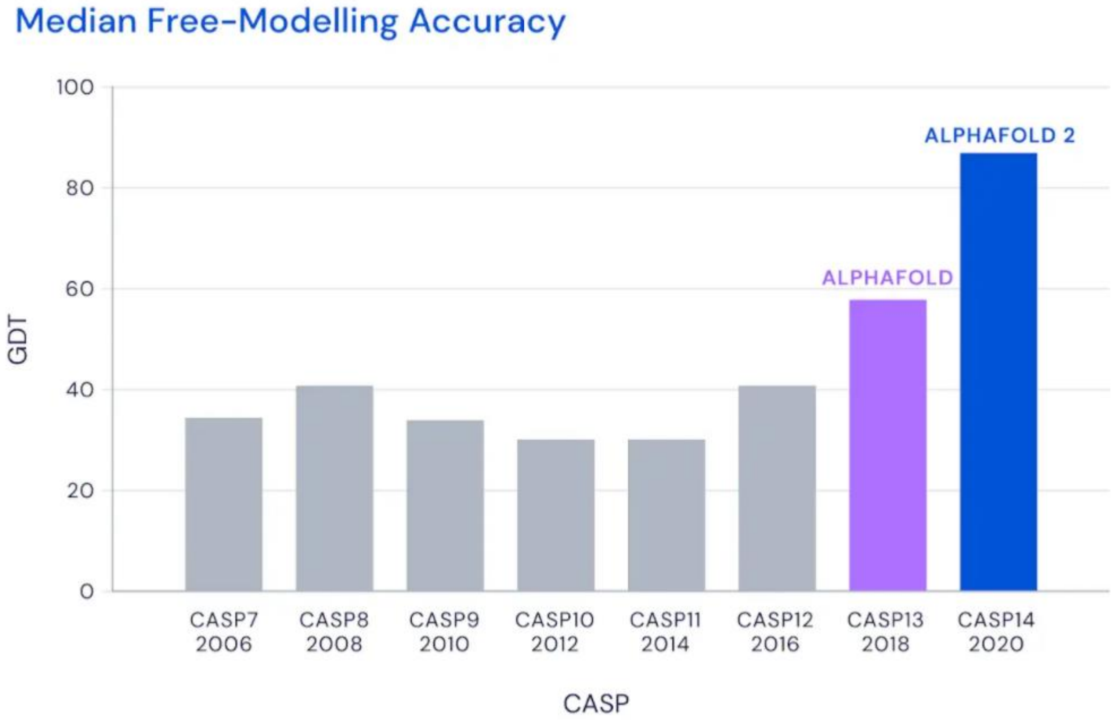
3D Protein Folding

- ***AlphaFold2*** has been acknowledged as the state of the art artificial intelligence system that solved this major challenge in the 14th CASP (Critical Assessment of protein Structure Prediction) competition.

- The metric used by CASP is GDT (Global Distance Test) that measures the percentage of amino acid residues that rely within a threshold distance from the correct position.

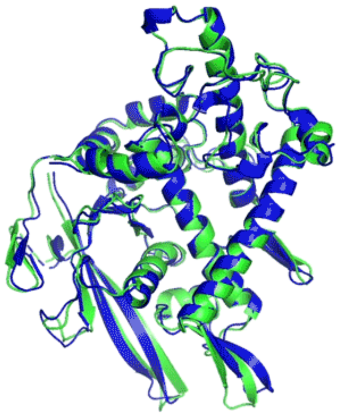
3D Protein Folding

- AlphaFold2 achieved a median score of 92.4 GDT totally for all targets.

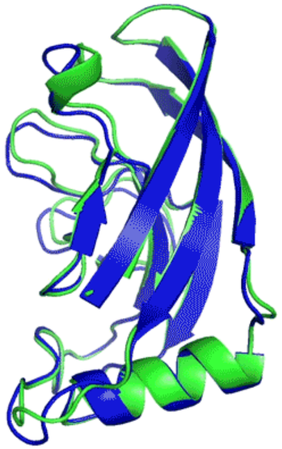


Reference: [AlphaFold]

3D Protein Folding



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



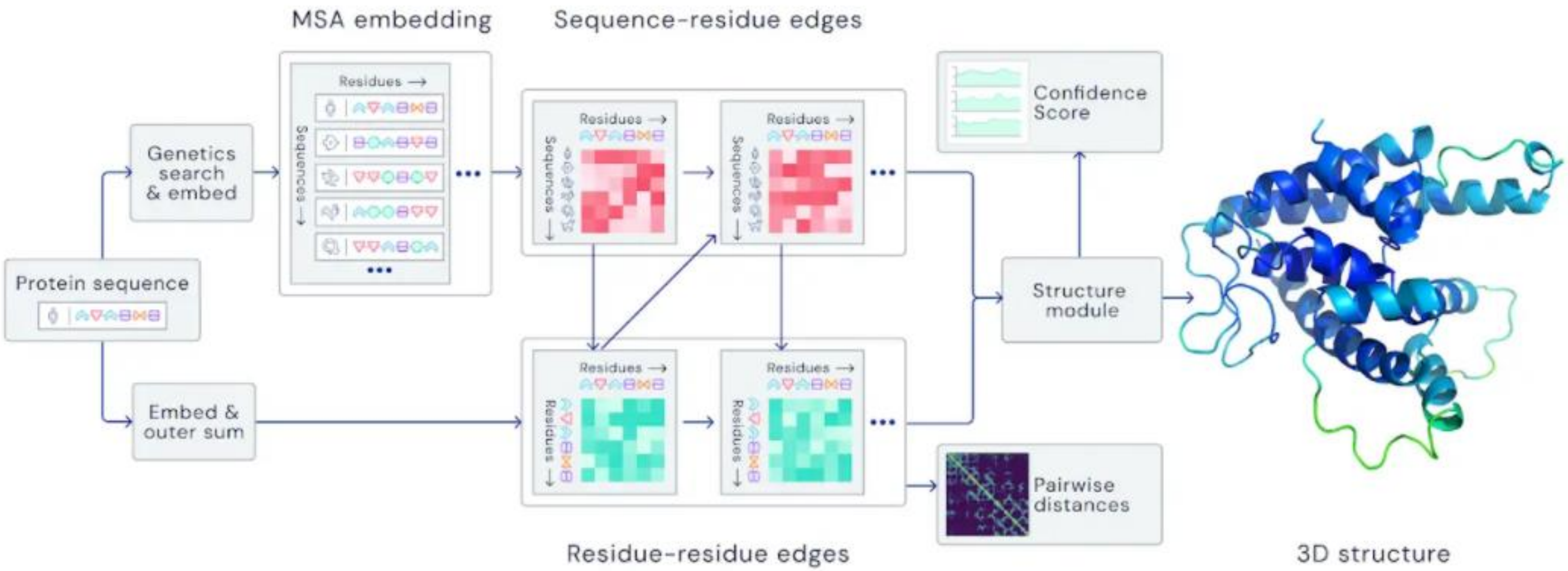
T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

Reference: [AlphaFold]

3D Protein Folding

- AlphaFold2 is an attention-based neural network system that is trained end-to-end.



Reference: [AlphaFold]

Bibliography

[DNA] https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

[GENE] https://socratic.org/questions/56116c7e11ef6b17032ddc8b

[RNA] https://byjus.com/biology/structure-of-rna/

[PROTEIN] https://en.wikipedia.org/wiki/Protein

[AMIN]https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_(Easlon)/Readings/04.3%3A_Amino_Acids

[KHAN]https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

[AlphaFold]https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr