

# **Mathematics of Bioinformatics**

## **---Theory, Practice, and Applications (Part II)**

**Matthew He, Ph.D.**

**Professor/Director**

**Division of Math, Science, and Technology**

**Nova Southeastern University, Florida, USA**

**December 18-21, 2010, Hong Kong, China**

**BIBM 2010**



# Part II Biological Functions, Networks, Systems Biology and Cognitive Informatics

---

---

## 6. Biological Networks and Graph Theory

- ❖ Introduction
- ❖ Graph Theory and Network Topology
- ❖ Models of Biological Networks
- ❖ Challenges and Perspectives



## 6.1. INTRODUCTION

- A mathematical graph is an abstract representation of a set of objects where some pairs of the objects are connected by links. Graph theory plays an important role in a wide variety of disciplines, ranging from communications to molecular and population biology.
- Network approaches offer the tools to analyze and understand a host of biological systems. In particular within the cell the variety of interactions between genes, proteins and metabolites are captured by network representations.



## 6.1. INTRODUCTION

- Biological networks are abstract representations of biological systems, which capture many of their essential characteristics.
- Biological systems ranging from food webs in ecology to biochemical interactions in molecular biology can be modeled and analyzed as networks.
- With the availability of complete genome sequences and high-throughput technologies and post-genomics experimental data, we have seen a growing interest in the study of networks of bio-molecular interactions in recent years.

## EXAMPLES OF NETWORKS AND REFERENCES

Graph	Nodes (vertices)	Links (edges)	Networks	References
Undirected graphs	Routers	Wires	Internet	Faloutsos, 1999
Directed graphs	Web pages	URL	World Wide Web networks	Albert, Jeong, Barabási, 1999
Directed graphs/ Undirected graphs	Genes	Expressions of gens A and B are correlated/ Regulatory influences	Gene regulatory networks	Lee et al., 2002
Directed graphs	Genes and Proteins	Transcription factor regulates a gene	Transcriptional regulatory networks	Guelzim et al., 2002
Directed bipartite graphs	Metabolites/ Reactions	Production/ Consumption	Metabolic networks	Savageau, 1991
Directed graphs	Proteins	Interaction	Protein interaction networks	Uetz et al., 2000
Directed graphs	People	Friendship/ Collaborations/ Sexual contacts/ Co-authorship of scientific papers	Societal networks	Milgram, 1967 Wasserman, 1994 Liljeros et al., 2001 Barabási et al., 2002



## 6.2. GRAPH THEORY AND NETWORK TOPOLOGY

- A graph is an ordered pair  $G := (V, E)$  comprising a set  $V$  of vertices or nodes together with a set  $E$  of edges or lines.
- The vertices belonging to an edge are called the ends, endpoints, or end vertices of the edge. A vertex may exist in a graph and not belong to an edge.
- $V$  and  $E$  are usually taken to be finite.
- The order of a graph is the number of vertices. A graph's size is the number of edges.
- The degree of a vertex is the number of edges that connect to it, where an edge that connects to the vertex at both ends (a loop) is counted twice.



## 6.2. GRAPH THEORY AND NETWORK TOPOLOGY

- Undirected graphs, if there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another.
- A directed graph or digraph is an ordered pair  $D := (V, A)$  is a set of ordered pairs of vertices and directed edges.
- A loop is an edge (directed or undirected) which starts and ends on the same vertex; these may be permitted or not permitted according to the application. In this context, an edge with two different ends is called a link.
- A graph is a weighted graph if a number (weight) is assigned to each edge. Such weights might represent, for example, costs, lengths or capacities, etc. depending on the problem.



## POWER LAW AND POWER LAW DISTRIBUTION

- Power laws are abundant in nature
- The power-law distribution has become the signature of biological networks
- A power law is any polynomial relationship that exhibits the property of scale invariance. The most common power laws relate two variables and have the form

$$P(x) = ax^k + o(x^k)$$

- where  $a$  and  $k$  are constants, and  $o(x^k)$  is an asymptotically small function of  $x$ . Here,  $k$  is typically called the scaling exponent, the word "scaling" denoting the fact that a power-law function satisfies  $P(cx) \propto P(x)$ , where  $c$  is a constant.



## POWER LAW AND POWER LAW DISTRIBUTION

- Logarithmic relation: a rescaling of the function's argument changes the constant of proportionality but preserves the shape of the function itself. This point becomes clearer if we take the logarithm of both sides:

$$\log P(x) = k \log x + \log a$$

- A few notable examples of power laws are the Gutenberg-Richter law for earthquake sizes, Pareto's law of income distribution, structural self-similarity of fractals, and scaling laws in biological systems.



## POWER LAW AND POWER LAW DISTRIBUTION

- A power-law distribution is any that, in the most general sense, has the form

$$P(x) \propto L(x)x^{-\alpha}$$

- where  $\alpha > 1$ , and  $L(x)$  is a slowly varying function, which is any function that satisfies  $L(tx)/L(x) \rightarrow 1$  as  $x \rightarrow \infty$  with  $t$  constant. This property of  $L(x)$  follows directly from the requirement that  $p(x)$  be asymptotically scale invariant; thus, the form of  $L(x)$  only controls the shape and finite extent of the lower tail.



## NETWORK TOPOLOGY AND NETWORK MODELS

- There are many tools and measures available now to study the structure and dynamics of complex networks.
- Statistical graph properties include the distribution of vertex degrees, the distribution of the clustering coefficients and other notions of density, the distribution of vertex-vertex distances, and the distribution of network motifs occurrences. In the following we will discuss three of the most fundamental quantities:
  - Degree distribution;
  - Clustering coefficient;
  - Centrality (degree, closeness, betweenness, and eigenvector) and essentiality.



## DEGREE DISTRIBUTION

- Degree (or connectivity) of a node in a network is the number of connections (edges) it has to other nodes. If a network is directed, then nodes have two different degrees, the in-degree, which is the number of incoming edges, and the out-degree, which is the number of outgoing edges.
- The degree distribution is the probability distribution of these degrees over the whole network. Formally, the degree distribution  $P(k)$  of a network is then defined to be the fraction of nodes in the network with degree  $k$ . Thus if there are  $n$  nodes in total in a network and  $n_k$  of them have degree  $k$ , we have  $P(k) = n_k/n$ .

$$P(k) = \frac{n_k}{n}$$



## DEGREE DISTRIBUTION

- The degree distribution is very important in studying biological networks and other complex networks. The simplest network model, for example, the (Bernoulli) random network, in which each of  $n$  nodes is connected (or not) with independent probability  $p$  (or  $1 - p$ ), has a binomial distribution of degrees

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

- or Poisson distribution in the limit of large  $n$ ,

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where  $\lambda$  is a constant.



## SCALE-FREE NETWORKS

- Most networks in the real world, however, have degree distributions very different from this.
- Most are highly right-skewed, meaning that a large majority of nodes have low degree, but a small number, known as "hubs", have high degree. Some networks, notably the Internet, the world-wide web, and some social networks are found to have degree distributions that approximately follow a power law:

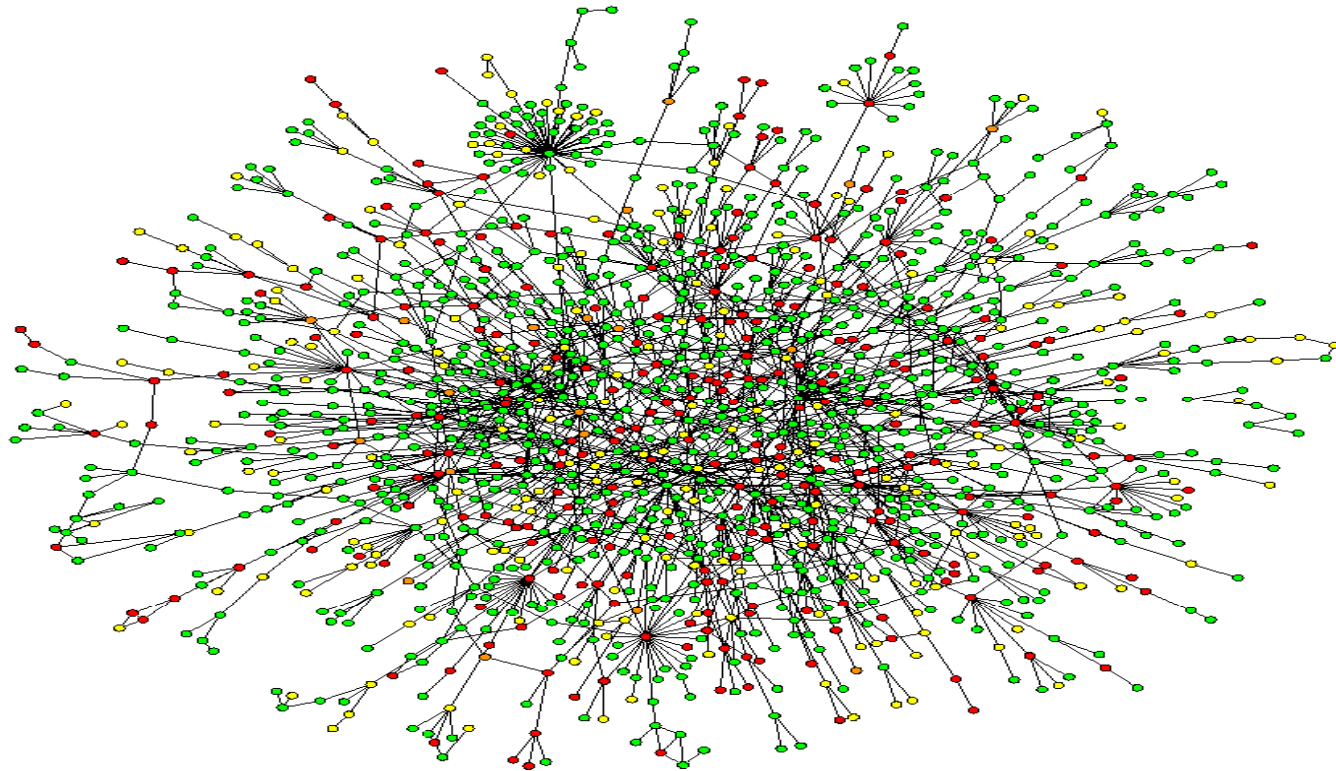
$$P(k) \approx k^{-\lambda}$$

- where  $\gamma$  is a constant. Such networks are called scale-free networks and have attracted particular attention for their structural and dynamical properties.

# BIOLOGICAL NETWORKS

Recent advances in biological and molecular networks presented  
a map of protein to-protein interactions

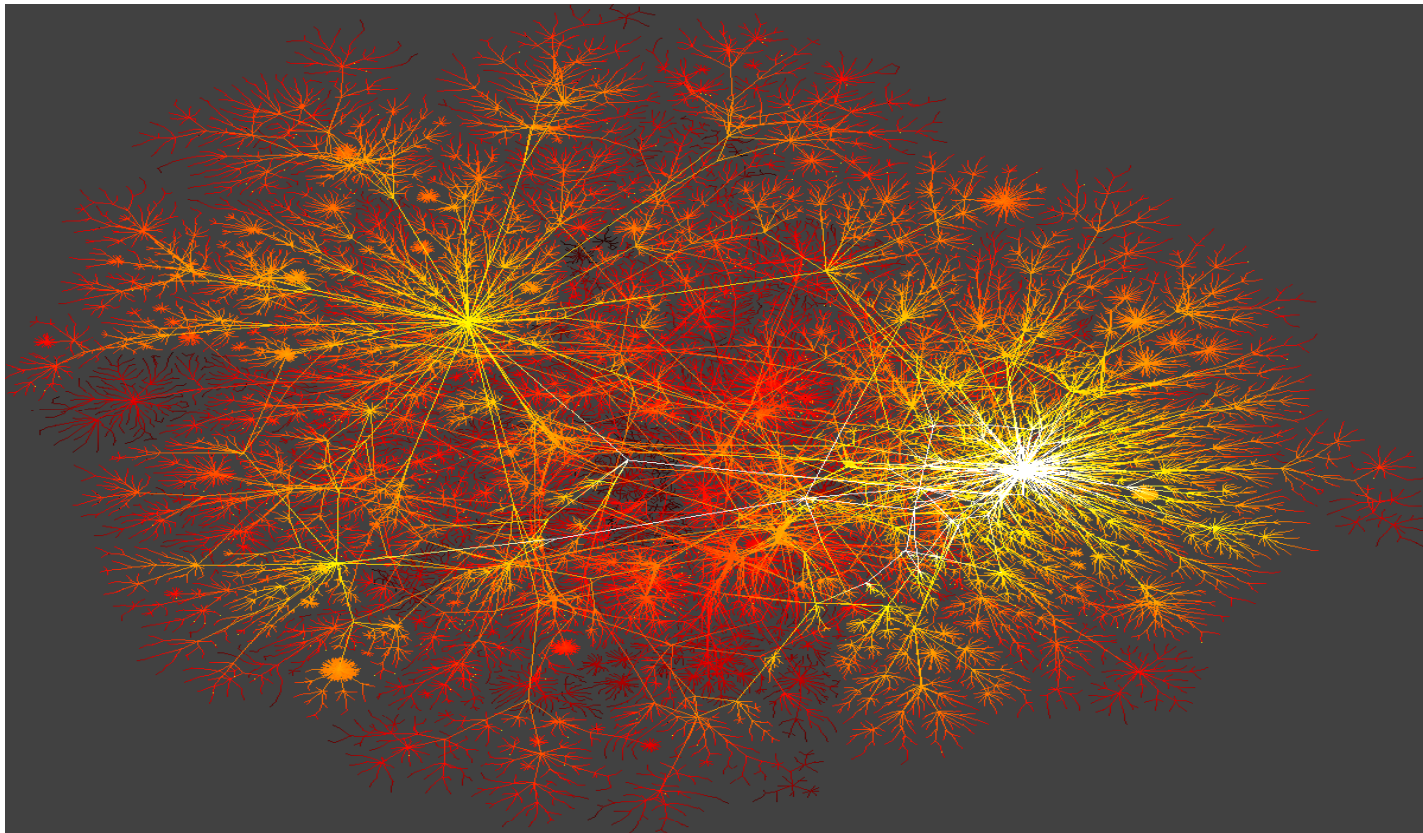
<http://www.nd.edu/~networks/Image/gallery.htm>



## WEB NETWORKS

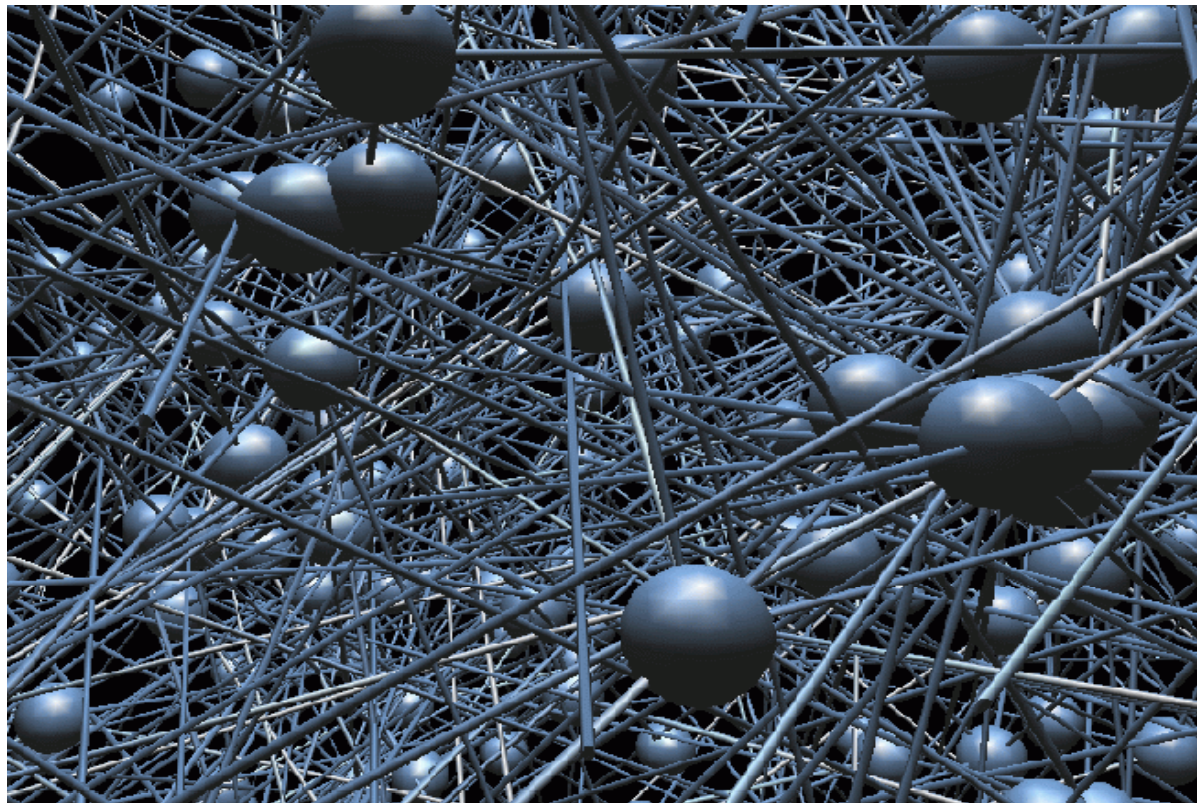
Internet connectivity was demonstrated as a scale free network.

<http://www.nd.edu/~networks/Image/gallery.htm>



## **SOCIAL NETWORKS**

Image of Social links in Canberra, Australia by A. S. Klovdahl at  
<http://www.nd.edu/~networks/Image/gallery.htm>





## CENTRALITY AND ESSENTIALITY

- Within graph theory and network analysis, there are various measures of the centrality of a vertex within a graph that determine the relative importance of a vertex within the graph.
- There are four measures of centrality that are widely used in network analysis:
  - degree centrality,
  - closeness,
  - betweenness,
  - eigenvector centrality.



## CENTRALITY AND ESSENTIALITY

- **Degree centrality** is the most basic of the centrality measures. It is defined as the number of links incident upon a node (i.e., the number of ties that a node has). Degree is often interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information). If the network is directed, then we usually define two separate measures of degree centrality, namely indegree and outdegree. Indegree is a count of the number of ties directed to the node, and outdegree is the number of ties that the node directs to others. For a graph  $G := (V, E)$  with  $n$  vertices, the degree centrality  $C_D(v)$  for vertex  $v$  is:

$$C_D(v) = \frac{\text{deg}(v)}{(n-1)}$$



## CENTRALITY AND ESSENTIALITY

- **Closeness** is a centrality measure of a vertex within a graph. Vertices that are 'shallow' to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. Closeness is preferred in network analysis to mean shortest-path length, as it gives higher values to more central vertices, and so is usually positively associated with other measures such as degree. It is defined as the mean geodesic distance (i.e the shortest path) between a vertex  $v$  and all other vertices reachable from it:

$$\frac{\sum_{t \in V \setminus v} d_G(v, t)}{(n - 1)}$$

where  $n > 1$  is the size of the network's 'connectivity component'  $V$  reachable from  $v$ . Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to other reachable vertices in the network.



## CENTRALITY AND ESSENTIALITY

- **Betweenness** is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not. For a graph  $G := (V, E)$  with  $n$  vertices, the betweenness  $C_B(v)$  for vertex  $v$  is

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where  $\sigma_{st}$  is the number of shortest geodesic paths from  $s$  to  $t$ , and  $\sigma_{st}(v)$  is the number of shortest geodesic paths from  $s$  to  $t$  that pass through a vertex  $v$ . This may be normalized by dividing through the number of pairs of vertices not including  $v$ , which is  $(n - 1)(n - 2)$ . Calculating the betweenness and closeness centralities of all the vertices in a graph involves calculating the shortest paths between all pairs of vertices on a graph.



## CENTRALITY AND ESSENTIALITY

- **Eigenvector centrality** is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.
- Let  $x_i$  denote the score of the  $i$ th node. Let  $A_{i,j}$  be the adjacency matrix of the network. Hence  $A_{i,j} = 1$  if the  $i$ th node is adjacent to the  $j$ th node, and  $A_{i,j} = 0$  otherwise. More generally, the entries in  $A$  can be real numbers representing connection strengths.
- For the  $i^{\text{th}}$  node, let the centrality score be proportional to the sum of the scores of all nodes which are connected to it. Hence

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^N A_{i,j} x_j$$



## CENTRALITY AND ESSENTIALITY

- where  $M(i)$  is the set of nodes that are connected to the  $i^{\text{th}}$  node,  $N$  is the total number of nodes and  $\lambda$  is a constant. In vector notation this can be rewritten as

$$X = \frac{1}{\lambda} AX, AX = \lambda X$$

- In general, there will be many different eigenvalues  $\lambda$  for which an eigenvector solution exists. However, the additional requirement that all the entries in the eigenvector be positive implies (by the Perron–Frobenius theorem) that only the greatest eigenvalue results in the desired centrality measure. The  $i^{\text{th}}$  component of the related eigenvector then gives the centrality score of the  $i^{\text{th}}$  node in the network. Power iteration is one of many eigenvalue algorithms that may be used to find this dominant eigenvector.



## THE NETWORK MODELS

- Random network model;
  - Scale-free network model;
  - Hierarchical network model.
- 
- **Random network:** A random network is obtained by starting with a set of  $n$  vertices and adding edges between them at random. Different random graph models produce different probability distributions on graphs. Most commonly studied is the Erdős–Rényi model (Erdős–Rényi, 1960) denoted  $G(n,p)$ , in which every possible edge occurs independently with probability  $p$ . This process generates a graph with approximately  $pN(N-1)/2$  randomly distributed edges.



## THE NETWORK MODELS

- **Scale-free network:** A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction  $P(k)$  of nodes in the network having  $k$  connections to other nodes goes for large values of  $k$  as  $P(k) \sim k^{-\gamma}$  where  $\gamma$  is a constant whose value is typically in the range  $2 < \gamma < 3$ , although occasionally it may lie outside these bounds. Scale-free networks are noteworthy because many empirically observed networks appear to be scale-free, including the world wide web, protein networks, citation networks, and some social networks.
- **Hierarchical network:** The type of network topology in which a central 'root' node (the top level of the hierarchy) is connected to one or more other nodes that are one level lower in the hierarchy (i.e., the second level) with a point-to-point link between each of the second level nodes and the top level central 'root' node. The hierarchy of the tree is symmetrical.



## 6.3. MODELS OF BIOLOGICAL NETWORKS

### GENE REGULATORY NETWORKS

- A gene regulatory network (GRN) or genetic regulatory network is a collection of DNA segments in a cell which interact with each other and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA. As we see here, a GRN involves interaction between, DNA, RNA, proteins and other molecules.
- A gene regulatory network can be viewed as a directed graph: a pair  $(V, E)$  where  $V$  is a set of vertices (genes) and  $E$  a set of directed edges (regulatory influences), pair  $(A, B)$  of vertices, where  $A$  is the source vertex and  $B$  the target vertex. A gene  $A$  directly regulates a gene  $B$  if the protein that is encoded by  $A$  is a transcription factor for gene  $B$ . This simple model can be improved by adding additional attribute on vertices or edges: for example, “+” or “-” labels on edges may indicate positive or negative regulatory influence.



## GENE REGULATORY NETWORKS

- Typically, a gene regulatory network is modeled as a system of rate equations, describing the reaction kinetics of the constituent parts and governing the evolution of mRNA and protein concentrations. Suppose that our regulatory network has  $N$  nodes, and let  $S_1(t), S_2(t), \dots, S_N(t)$  represent the concentrations of the  $N$  corresponding substances at time  $t$ . Then the temporal evolution of the system can be described approximately by

$$\frac{dS_j}{dt} = f_j(S_1(t), S_2(t), \dots, S_N(t))$$

where the functions  $f_j$  express the dependence of  $S_j$  on the concentrations of other substances present in the cell. By solving for the fixed point of the system:

$$\frac{dS_j}{dt} = 0$$

for all  $j$ , one obtains (possibly several) concentration profiles of proteins and mRNAs that are theoretically sustainable (though not necessarily stable).



## **PROTEIN INTERACTION NETWORKS**

- A key feature of the biological organization in all organisms is the tendency of proteins with a common function to physically associate via stable protein to protein interactions (PPI) to form larger macromolecular assemblies.
- These protein complexes are often linked together by extended networks of weaker, transient PPI, to form interaction networks that integrate pathways mediating the major cellular processes.
- The analysis of protein interaction networks is an important and very active research area in bioinformatics and computational biology.



## METABOLIC NETWORKS

- Metabolic networks comprise the chemical reactions of metabolism as well as the regulatory interactions that guide these reactions.
- With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to humans.
- Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG, a database resource for linking genome to life and the environment. Since 1995 KEGG has been developed by the Kanehisa Laboratories in the Kyoto University Bioinformatics Center and the Human Genome Center of the University of Tokyo as part of their research activities.), EcoCyc (EcoCyc is a scientific database for the bacterium *Escherichia coli* K-12 MG1655).



## 6.4. CHALLENGES AND PERSPECTIVES

- A major challenge consists in identifying with reasonable accuracy these complex molecular interactions that take place at different levels from genes to metabolites through proteins.
- Analysis will require development of appropriate computer data bases and development of new theory and algorithms in the mathematical theory of directed graphs.
- Understanding the evolution of such genomic networks under the influence of point and chromosomal mutations that literally scramble the genomic wiring diagram will require new uses of random directed graph theory, stochastic processes, and population genetic models.



# **Part II Biological Functions, Networks, Systems Biology and Cognitive Informatics**

---

---

## **7. Biological Systems, Fractals, and Systems Biology**

- ❖ Introduction
- ❖ Fractal Geometry Preliminaries
- ❖ Fractal Geometry in Biological Systems
- ❖ Systems Biology and Perspectives
- ❖ Challenges and Perspectives



## 7.1 INTRODUCTION

- To understand complex biological systems at the different levels of molecules, cells, tissues, and organs, etc., we must examine their structures, dynamics, and functions.
- It requires the integration of experimental, computational, and theoretical explorations.
- The study of biological systems cannot be limited to simply listing its parts (proteins, genes, cells, etc.). A deeper understanding of biological systems can demonstrate how these parts are assembled together and how they interact with each other and with the surrounding environment.



## 7.1 INTRODUCTION

- Contemporary science should recognize the importance of “wholeness”. Wholeness is defined as “problems of organization, phenomena not resolvable into local events, dynamic interactions manifest in the difference of behavior of parts when isolated or in higher configuration, etc.
- Biologists have traditionally modeled nature using Euclidean representations of natural objects or series. They represented heartbeats as sine waves, conifer trees as cones, animal habitats as simple areas, and cell membranes as curves or simple surfaces.



## 7.1 INTRODUCTION

- Scientists discovered that the basic architecture of a chromosome is tree-like; every chromosome consists of many 'mini-chromosomes', and therefore can be treated as fractal. For a human chromosome, for example, a fractal dimension  $D$  equals 2.34 (between the plane and the space dimension).
- Self-similarity has been found also in DNA sequences. In the opinion of some biologists, fractal properties of DNA can be used to resolve evolutionary relationships in animals.



## 7.2 FRACTAL GEOMETRY PRELIMINARIES

- Fractal geometry is a new language used to describe, model and analyze complex forms found in nature.
- A fractal is a geometric shape that has two most important properties:
  - The object is self-similar.
  - The object has fractional dimension.
- Iterated Function System (IFS) fractals are created on the basis of simple plane transformations: scaling, dislocation and the plane axes rotation.

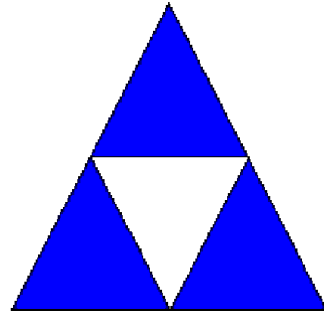


## 7.2 FRACTAL GEOMETRY PRELIMINARIES

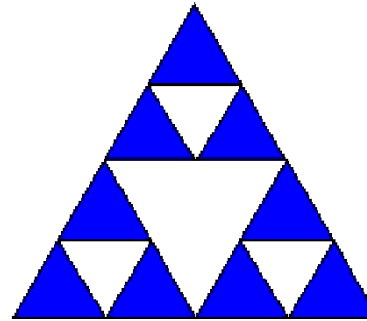
- Creating an IFS fractal consists of the following steps:
  - defining a set of plane transformations,
  - drawing an initial pattern on the plane (any pattern),
  - transforming the initial pattern using the transformations defined in the first step,
  - transforming the new picture (combination of initial and transformed patterns) using the same set of transformations,
  - repeating the fourth step as many times as possible (in theory, this procedure can be repeated an infinite number of times).



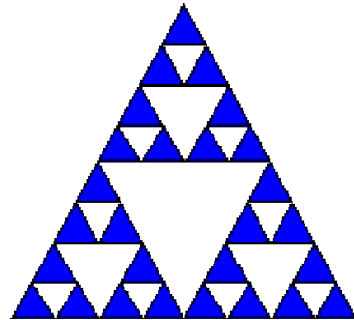
1)



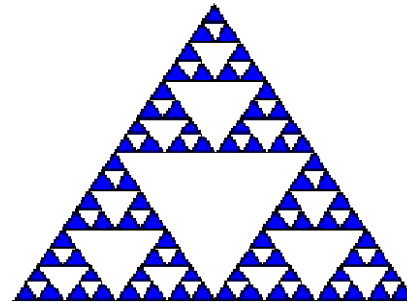
2)



3)



4)



## Sierpinski Triangles

The most famous ISF fractals are the Sierpinski Triangles



## FRACTAL DIMENSION

- Fractal dimension is a measure of how "complicated" a self-similar figure is. In a rough sense, it measures "how many points" lie in a given set. A plane is "larger" than a line, while  $S$  sits somewhere in between these two sets.
- The most commonly used methods for determining the fractal dimensions have been using the scaling relationship and the capacity dimension done by box counting. Examples where fractal dimension has been measured include the surfaces of proteins, cell membranes, cells of the cornea, and bacterial colonies.



## FRACTAL DIMENSION

- Using this fractal as an example, we can prove that the fractal dimension is not an integer.
- First of all we have to find out how the "size" of an object behaves when its linear dimension increases. In one dimension we can consider a line segment. If the linear dimension of the line segment is doubled, then the length (characteristic size) of the line has doubled also. In two dimensions, if the linear dimensions of a square for example is doubled then the characteristic size, the area, increases by a factor of 4. In three dimensions, if the linear dimension of a box is doubled then the volume increases by a factor of 8.



## FRACTAL DIMENSION

- For the square, we have  $N^2$  self-similar pieces, each with magnification factor  $N$ . So we can write

$$D = \frac{\log(\#self - similar)}{\log(magnif.factor)} = \frac{\log N^2}{\log N} = 2$$

- Similarly, the dimension of a cube is

$$D = \frac{\log(\#self - similar)}{\log(magnif.factor)} = \frac{\log N^3}{\log N} = 3$$



## FRACTAL DIMENSION

- Thus, we take as the *definition* of the fractal dimension of a self-similar object

$$D = \frac{\log(\#self - similar)}{\log(magnif . factor)}$$

- Now we can compute the dimension of **S**. For the Sierpinski triangle consists of 3 self-similar pieces, each with magnification factor 2. So the fractal dimension is

$$D = \frac{\log(\#self - similar)}{\log(magnif . factor)} = \frac{\log 3}{\log 2} \approx 1.58$$



## 7.3 FRACTAL GEOMETRY IN BIOLOGICAL SYSTEMS

- Fractal geometry reveals the regularity behind matter with apparently irregular forms.
- A fractal implies a complex pattern with self-similarity and self-affinity, i.e. a fractal has a shape made of parts similar to the whole in some way.
- DNA sequences and the structures of protein have such a complex form. Scaling behavior can be seen in fundamental biological structures.
- DNA structure demonstrates fractal properties in the distribution of sequence information.
- The fractal research into proteins and enzymes is currently an active field of bioinformatics.



## FRACTALS IN DNA SEQUENCES

- Self-similarity has recently been found in DNA sequences by Stanley 1992; Nonnenmacher et al. 1994.
- The multifractal spectrum approach has been used to reconstruct the evolutionary history of organisms from m-DNA sequences by Glazier, et al. 1995. The multifractal spectra for invertebrates and vertebrates were quite different, allowing for the recognition of broad groups of organisms. They concluded that DNA sequences display fractal properties, and that these can be used to resolve evolutionary relationships in animals.
- Furthermore, Xiao et al. (1995) found that nucleotide sequences in animals, plants and humans display fractal properties. They also showed that exon and intron sequences differ in their fractal properties



## FRACTALS IN CELL, PROTEIN AND CHROMOSOME STRUCTURES

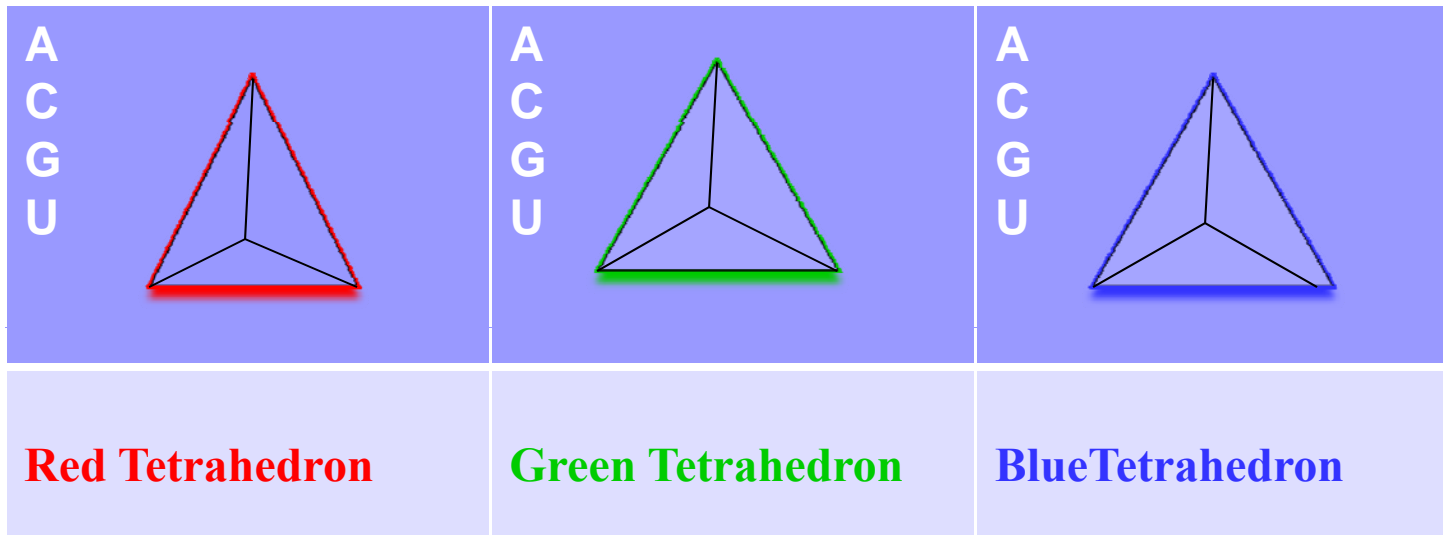
- Takahashi (1989) hypothesized that the basic architecture of a chromosome is tree-like, consisting of a concatenation of 'mini-chromosomes'. A fractal dimension of  $D = 2.34$  was determined from an analysis of first and second order branching patterns in a human metaphase chromosome.
- Xu et al. (1994) hypothesized that the twistings of DNA binding proteins have fractal properties.
- Lewis and Rees (1985) determined the fractal dimension of protein surfaces ( $2 \leq D \leq 3$ ) using microprobes. A mean surface dimension of  $D = 2.4$  was determined using microprobe radii ranging from 1-3.5 angstroms.
- More highly irregular surfaces ( $D > 2.4$ ) were found to be sites of inter-protein interaction.



## FRACTALS IN CELL, PROTEIN AND CHROMOSOME STRUCTURES

- Smith et al. (1989) used fractal dimension as a measure of contour complexity in two-dimensional images of neural cells. They recommend  $D$  as a quantitative morphological measure of cellular complexity.
- **Fractals and DNA Walk:** A DNA walk of a genome represents how the frequency of each nucleotide of a pairing nucleotide couple changes locally. This analysis implies measurement of the local distribution of Gs in the content of GC and of Ts in the content of TA. Lobry was the first to propose this analysis (1996, 1999).
- **RNA Tetrahedron:** A regular tetrahedron has four equal faces, four vertices and six edges (Minimal faces required to form a three dimensional polyhedron; similar to the fact that a triangle is the first polygon in two dimensional space). RNA bases consists of four bases A, C, G, U. We label each letter A, C, G, U to each face of a regular tetrahedron.

# FRACTALS IN CELL, PROTEIN AND CHROMOSOME STRUCTURES



- Biperiodic table of the genetic code, we roll three tetrahedrons and record three letters covered at the bottom of each toss. Assume that each event is equally likely. It's easy to see that there is a total of  $4 \times 4 \times 4 = 4^3 = 64$  possible outcomes. We list all these 64 elements in the following table  $G(i, j)$ ,  $i, j = 1, 2, 3, 4, 5, 6, 7, 8$ . Theoretically for each cell of the table, there are 64 possible ways to arrange a codon. The total number of codons in the table is  $64!$ .

**BIPERIODIC TABLE OF THE GENETIC CODE:  
FROM A FULL 4-ARY TREE (the height of the tree is 3)**

<b>AAA</b>	<b>ACA</b>	<b>AGA</b>	<b>AUA</b>	<b>CAA</b>	<b>CCA</b>	<b>CGA</b>	<b>CUA</b>
<b>AAC</b>	<b>ACC</b>	<b>AGC</b>	<b>AUC</b>	<b>CAC</b>	<b>CCC</b>	<b>CGC</b>	<b>CUC</b>
<b>AAG</b>	<b>ACG</b>	<b>AGG</b>	<b>AUG</b>	<b>CAG</b>	<b>CCG</b>	<b>CGG</b>	<b>CUG</b>
<b>AAU</b>	<b>ACU</b>	<b>AGU</b>	<b>AUU</b>	<b>CAU</b>	<b>CCU</b>	<b>CGU</b>	<b>CUU</b>
<b>GAA</b>	<b>GCA</b>	<b>GGA</b>	<b>GUA</b>	<b>UAA</b>	<b>UCA</b>	<b>UGA</b>	<b>UUA</b>
<b>GAC</b>	<b>GCC</b>	<b>GGC</b>	<b>GUC</b>	<b>UAC</b>	<b>UCC</b>	<b>UGC</b>	<b>UUC</b>
<b>GAG</b>	<b>GCG</b>	<b>GGG</b>	<b>GUG</b>	<b>UAG</b>	<b>UCG</b>	<b>UGG</b>	<b>UUG</b>
<b>GAU</b>	<b>GCU</b>	<b>GGU</b>	<b>GUU</b>	<b>UAU</b>	<b>UCU</b>	<b>UGU</b>	<b>UUU</b>



## FRACTAL PROPERTIES OF PROTEINS AND POLYMERS

- A polymer is a molecule that is composed of a series of "building blocks" (called monomers) connected to one another in a chain.
- Monomers are not connected in a straight line. Instead, the angles between the monomers can be different and the entire molecule can twist into pretty complicated shapes.
- The same is true for proteins, which are formed by amino acid bonding together in a chain. Twisting, as well as folding and breaking often implies by itself that the shape is fractal. Proteins and many other polymers are, indeed, fractal and various methods exist for finding their fractal dimension.



## FRACTAL PROPERTIES OF PROTEINS AND POLYMERS

- For some interesting proteins the results are shown in Table 7.10 (Ideker T. et al., 2001). Note that the dimensions are much higher than 1, which you would expect from a linear chain. This is another proof that proteins are fractal.

**TABLE 7.10 Fractal dimensions of some proteins**

Protein	Fractal Dimension
Lysozyme (egg-white)	1.614
Hemoglobin (oxygen carrier in the blood)	1.583
Myoglobin (muscle protein)	1.728



## FRACTAL SURFACES IN BIOLOGICAL SYSTEMS

- Fractal surfaces can be used to characterize the roughness or irregularity of protein surfaces (Lewis and Rees, 1985).
- The degree of irregularity of a surface may be described by the fractal dimension  $D$ . For protein surfaces defined with probes in the range of 1.0 to 3.5 angstroms in radius,  $D$  is approximately 2.4 or intermediate between the value for a completely smooth surface ( $D = 2$ ) and that for a completely space-filling surface ( $D = 3$ ).
- Individual regions of proteins show considerable variation in  $D$ . These variations may be related to structural features such as active sites and subunit interfaces, suggesting that surface texture may be a factor influencing molecular interactions.



## 7.4 SYSTEMS BIOLOGY AND PERSPECTIVES

- The emerging field of systems biology involves the application of experimental, theoretical, and modeling techniques to the study of biological organisms at all levels, from the molecular, through the cellular, to the behavioral. Its aim is to understand biological processes as whole systems instead of as isolated parts.
- The overall systems biology methodology includes the formulation of a model once the components of the system have been defined, followed by the systematical perturbation (either genetically or environmentally) and monitoring of the system. The experimentally observed responses are then reconciled with those predicted by the model.
- To understand biology at the system level, we must examine the structure and dynamics of cellular function, rather than the characteristics of isolated parts of a cell or organism. Properties of systems, such as robustness, emerge as central issues, and understanding these properties may have an impact on the future of medicine.



## 7.5 CHALLENGES AND PERSPECTIVES

- The analysis of complex hierarchical systems therefore represents one of the most important open areas in biology.
- The organization and integration of these details into a functional biological system will require the techniques of the mathematician as well as the data of the biologist. Problems of this sort are at the core of genetics, neurobiology, developmental biology and immunology.
- The central theoretical question is how the molecular details are integrated into a functional unity, a question central to at least three major fields: neurobiology, developmental biology, and immunology.
- Applications in biology require the development of effective computational methods for the analysis of dynamical systems and their bifurcations. New mathematics is emerging from work in this direction.



## **Part II Biological Functions, Networks, Systems Biology and Cognitive Informatics**

---

---

### **8. Matrix Genetics, Hadamard Matrix, and Algebraic Biology**

- ❖ Introduction
- ❖ Genetic Matrices and the Degeneracy of the Genetic Code
- ❖ The Genetic Code and Hadamard Matrices
- ❖ Genetic Yin-Yang Algebras
- ❖ Challenges and Perspectives



## 8.1 INTRODUCTION

- Science has led to a new understanding of life itself: «**Life is a partnership between genes and mathematics**» (Stewart I. Life's other secret: The new mathematics of the living world. 1999, New-York: Penguin).
- But what kind of mathematics is a partner with the genetic code and defines the structure of living matter? This talk presents some results of matrix approaches to the genetic code to investigate such a partner: “genetic mathematics”.
- We will show a close phenomenological connection of the genetic code in matrix forms of its presentation with Rademacher functions, Walsh functions and Hadamard matrices, which are known in the theory of noise-immunity coding, spectral analysis and quantum computers. These materials are important for development of algebraic biology.



In computers, information is stored in a form of matrices. Noise-immunity coding uses matrices also, for example, Kronecker families of Hadamard matrices:

$$H_2 = \begin{array}{|c|c|} \hline 1 & 1 \\ \hline -1 & 1 \\ \hline \end{array} ; H_4 = \begin{array}{|c|c|c|c|} \hline 1 & 1 & 1 & 1 \\ \hline -1 & 1 & -1 & 1 \\ \hline -1 & -1 & 1 & 1 \\ \hline 1 & -1 & -1 & 1 \\ \hline \end{array} ; H_8 = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ \hline -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ \hline 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ \hline -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ \hline 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ \hline 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ \hline -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ \hline \end{array}$$

Here black cells correspond to components “+1”, white cells – to “-1”.

## 8.2 DEGENERACY OF THE GENETIC CODE

By analogy we study the Kronecker family of symbolic matrices  $[C \ A; U \ G]^{(n)}$  of the genetic alphabet  $A, C, G, U/T$  (adenine, cytosine, guanine, uracil/thymine).

Here each matrix

$[C \ A; U \ G]^{(n)}$

contains

the complete

set of n-plets

in a strong order.


$$[C \ A; U \ G] = \begin{bmatrix} C & A \\ U & G \end{bmatrix}; [C \ A; U \ G]^{(2)} = \begin{bmatrix} CC & CA & AC & AA \\ CU & CG & AU & AG \\ UC & UA & GC & GA \\ UU & UG & GU & GG \end{bmatrix}$$

$$[C \ A; U \ G]^{(3)} = \begin{bmatrix} CCC & CCA & CAC & CAA & ACC & ACA & AAC & AAA \\ CCU & CCG & CAU & CAG & ACU & ACG & AAU & AAG \\ CUC & CUA & CGC & CGA & AUC & AUA & AGC & AGA \\ CUU & CUG & CGU & CGG & AUU & AUG & AGU & AGG \\ UCC & UCA & UAC & UAA & GCC & GCA & GAC & GAA \\ UCU & UCG & UAU & UAG & GCU & GCG & GAU & GAG \\ UUC & UUA & UGC & UGA & GUC & GUA & GGC & GGA \\ UUU & UUG & UGU & UGG & GUU & GUG & GGU & GGG \end{bmatrix}$$

$$[C A; U G] = \begin{array}{|c|c|} \hline C & A \\ \hline U & G \\ \hline \end{array}; [C A; U G]^{(2)} = \begin{array}{|c|c|c|c|} \hline CC & CA & AC & AA \\ \hline CU & CG & AU & AG \\ \hline UC & UA & GC & GA \\ \hline UU & UG & GU & GG \\ \hline \end{array}$$

$$[C A; U G]^{(3)} = \begin{array}{|c|c|c|c|c|c|c|c|} \hline CCC & CCA & CAC & CAA & ACC & ACA & AAC & AAA \\ \hline CCU & CCG & CAU & CAG & ACU & ACG & AAU & AAG \\ \hline CUC & CUA & CGC & CGA & AUC & AUA & AGC & AGA \\ \hline CUU & CUG & CGU & CGG & AUU & AUG & AGU & AGG \\ \hline UCC & UCA & UAC & UAA & GCC & GCA & GAC & GAA \\ \hline UCU & UCG & UAU & UAG & GCU & GCG & GAU & GAG \\ \hline UUC & UUA & UGC & UGA & GUC & GUA & GGC & GGA \\ \hline UUU & UUG & UGU & UGG & GUU & GUG & GGU & GGG \\ \hline \end{array}$$

For example, the genetic matrix  $[C A; U G]^{(3)}$  contains all 64 triplets which encode 20 amino acids and punctuation signals in all biological organisms. We will show that this genomatrix  $[C A; U G]^{(3)}$  has unexpected algebraic connection with a specificity of the degeneracy of the genetic code.

- 
- It is well known that 17 variants (or dialects) of the genetic code which differ one from another by some details of correspondences between triplets and objects encoded by them (<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi> ).

Most of these dialects have a general scheme of degeneracy of the genetic code (including main dialects such as the Standard Code and the Vertebrate Mitochondrial Code which is the most symmetrical) .



In this basic scheme of the degeneracy, the set of 64 triplets is divided into two equal subsets with 32 “black” triplets and 32 “white” triplets.

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

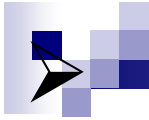
### THE STANDARD CODE

8 subfamilies of the “two-position NN-triplets” (“black triplets”) and the amino acids, which are encoded by them	8 subfamilies of the “three-position NN-triplets” („white triplets”) and the amino acids, which are encoded by them
<u>CCC</u> , <u>CCU</u> , <u>CCA</u> , <u>CCG</u> → Pro	<u>CAC</u> , <u>CAU</u> , <u>CAA</u> , <u>CAG</u> → His, His, Gln, Gln
<u>CUC</u> , <u>CUU</u> , <u>CUA</u> , <u>CCG</u> → Leu	<u>AAC</u> , <u>AAU</u> , <u>AAA</u> , <u>AAG</u> → Asn, Asn, Lys, Lys
<u>CGC</u> , <u>CGU</u> , <u>CGA</u> , <u>CGG</u> → Arg	<u>AUC</u> , <u>AUU</u> , <u>AUA</u> , <u>AUG</u> → Ile, Ile, Ile, Met
<u>ACC</u> , <u>ACU</u> , <u>ACA</u> , <u>ACG</u> → Thr	<u>AGC</u> , <u>AGU</u> , <u>AGA</u> , <u>AGG</u> → Ser, Ser, Arg, Arg
<u>UCC</u> , <u>UCU</u> , <u>UCA</u> , <u>UCG</u> → Ser	<u>UAC</u> , <u>UAU</u> , <u>UAA</u> , <u>UAG</u> → Tyr, Tyr, Stop, Stop
<u>GCC</u> , <u>GCU</u> , <u>GCA</u> , <u>GCG</u> → Ala	<u>UUC</u> , <u>UUU</u> , <u>UUA</u> , <u>UUG</u> → Phe, Phe, Leu, Leu
<u>GUC</u> , <u>GUU</u> , <u>GUA</u> , <u>GUG</u> → Val	<u>UGC</u> , <u>UGU</u> , <u>UGA</u> , <u>UGG</u> → Cys, Cys, Stop, Trp
<u>GGC</u> , <u>GGU</u> , <u>GGA</u> , <u>GGG</u> → Gly	<u>GAC</u> , <u>GAU</u> , <u>GAA</u> , <u>GAG</u> → Asp, Asp, Glu, Glu

### THE VERTEBRATE MITOCHONDRIAL CODE

8 subfamilies of the “two-position NN-triplets” (“black triplets”) and the amino acids, which are encoded by them	8 subfamilies of the “three-position NN-triplets” („white triplets”) and the amino acids, which are encoded by them
<u>CCC</u> , <u>CCU</u> , <u>CCA</u> , <u>CCG</u> → Pro	<u>CAC</u> , <u>CAU</u> , <u>CAA</u> , <u>CAG</u> → His, His, Gln, Gln
<u>CUC</u> , <u>CUU</u> , <u>CUA</u> , <u>CUG</u> → Leu	<u>AAC</u> , <u>AAU</u> , <u>AAA</u> , <u>AAG</u> → Asn, Asn, Lys, Lys
<u>CGC</u> , <u>CGU</u> , <u>CGA</u> , <u>CGG</u> → Arg	<u>AUC</u> , <u>AUU</u> , <u>AUA</u> , <u>AUG</u> → Ile, Ile, Met, Met
<u>ACC</u> , <u>ACU</u> , <u>ACA</u> , <u>ACG</u> → Thr	<u>AGC</u> , <u>AGU</u> , <u>AGA</u> , <u>AGG</u> → Ser, Ser, Stop, Stop
<u>UCC</u> , <u>UCU</u> , <u>UCA</u> , <u>UCG</u> → Ser	<u>UAC</u> , <u>UAU</u> , <u>UAA</u> , <u>UAG</u> → Tyr, Tyr, Stop, Stop
<u>GCC</u> , <u>GCU</u> , <u>GCA</u> , <u>GCG</u> → Ala	<u>UUC</u> , <u>UUU</u> , <u>UUA</u> , <u>UUG</u> → Phe, Phe, Leu, Leu
<u>GUC</u> , <u>GUU</u> , <u>GUA</u> , <u>GUG</u> → Val	<u>UGC</u> , <u>UGU</u> , <u>UGA</u> , <u>UGG</u> → Cys, Cys, Trp, Trp
<u>GGC</u> , <u>GGU</u> , <u>GGA</u> , <u>GGG</u> → Gly	<u>GAC</u> , <u>GAU</u> , <u>GAA</u> , <u>GAG</u> → Asp, Asp, Glu, Glu

The schemes of degeneracy for the Standard Code and the Vertebrate Mitochondrial Code with 32 “black” triplets and 32 “white” triplets.



The first subset contains those 32 “black” triplets, coding values of which are independent of a letter on their third position. For example, each of the four triplets with identical two first letters CGC, CGA, CGU, CGG encodes the same amino acid Arg. All such a type of NN-triplets we mark as “black triplets”.

The second subset contains those 32 “white” triplets, coding values of which depend on a letter on their third position. For example in the family of triplets with identical two first letters CAC, CAA, CAU, CAC, two triplets CAC, CAU encode amino acid His and the other two CAA, CAG encode another amino acid Gln. We mark all such a type of NN-triplets as “white triplets”.

- The unexpected **phenomenological** fact is a symmetrical disposition of black triplets and white triplets in the genomatrix  $[C\ A; U\ G]^{(3)}$ , which was constructed formally without any mention about amino acids and the degeneracy of the genetic code:

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

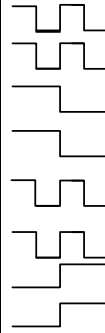


CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

Symmetrical properties of the genomatrix [C A; U G]<sup>(3)</sup>:

- 1) The left and right halves of the matrix mosaic are mirror-anti-symmetric to each other in its colors: any pair of cells, disposed by mirror-symmetrical manner in the halves, possesses the opposite colors;
- 2) Both quadrants along each diagonals are identical from the viewpoint of their mosaic;

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG



→  $B_{123} =$

1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1
1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1
1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1
-1	-1	-1	-1	1	1	1	1
-1	-1	-1	-1	1	1	1	1

3) Mosaics of all rows have a meander-line character, which is identical to **Rademacher functions** which are famous in the theory of discrete signals processing, spectral analysis and probability theory and which contain components “+1” and “-1” only.

Each row presents one of the Rademacher functions  $r_n(t) = \text{sign}(\sin 2^n \pi t)$ ,  $n = 1, 2, 3, \dots$ , if each black (white) cell is interpreted as it contains number “+1” (“-1”).



Digital signals processing pays a special attention to permutations of elements. By analogy let us study all six possible variants of permutations of positions inside triplets (1-2-3, 2-3-1, 3-1-2, 3-2-1, 2-1-3, 1-3-2). For example in the case of the cyclic shift of positions 1-2-3  $\rightarrow$  2-3-1 inside all triplets, a new genomatrix arises, where, for example, the black triplet CGA is replaced by the white triplet GAC, etc.:

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

CCC	CAC	ACC	AAC	CCA	CAA	ACA	AAA
CUC	CGC	AUC	AGC	CUA	CGA	AUA	AGA
UCC	UAC	GCC	GAC	UCA	UAA	GCA	GAA
UUC	UGC	GUC	GGC	UUA	UGA	GUA	GGA
CCU	CAU	ACU	AAU	CCG	CAG	ACG	AAG
CUU	CGU	AUU	AGU	CUG	CGG	AUG	AGG
UCU	UAU	GCU	GAU	UCG	UAG	GCG	GAG
UUU	UGU	GUU	GGU	UUG	UGG	GUG	GGG



The 6 different genomatrices exist for all possible 6 variants of positional permutations inside triplets

(1-2-3, 2-3-1, 3-1-2, 3-2-1, 2-1-3, 1-3-2):

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

CCC	CAC	ACC	AAC	CCA	CAA	ACA	AAA
CUC	CGC	AUC	AGC	CUA	CGA	AUA	AGA
UCC	UAC	GCC	GAC	UCA	UAA	GCA	GAA
UUC	UGC	GUC	GGC	UUA	UGA	GUA	GGA
CCU	CAU	ACU	AAU	CCG	CAG	ACG	AAG
CUU	CGU	AUU	AGU	CUG	CGG	AUG	AGG
UCU	UAU	GCU	GAU	UCG	UAG	GCG	GAG
UUU	UGU	GUU	GGU	UUG	UGG	GUG	GGG

CC	CCA	ACC	ACA	CAC	CAA	AAC	AAA
CCU	CCG	ACU	ACG	CAU	CAG	AAU	AAG
UCC	UCA	GCC	GCA	UAC	UAA	GAC	GAA
UCU	UCG	GCU	GCG	UAU	UAG	GAU	GAG
CUC	CUA	AUC	AUA	CGC	CGA	AGC	AGA
CUU	CUG	AUU	AUG	CGU	CGG	AGU	AGC
UUC	UUA	GUC	GUA	UGC	UGA	GGC	GGA
UUU	UUG	GUU	GUG	UGU	UGG	GGU	GGG

CCC	ACC	CAC	AAC	CCA	ACA	CAA	AAA
UCC	GCC	UAC	GAC	UCA	GCA	UAA	GAA
CUC	AUC	CGC	AGC	CUA	AUA	CGA	AGA
UUC	GUC	UGC	GGC	UUA	GUA	UGA	GGA
CCU	ACU	CAU	AAU	CCG	ACG	CAG	AAG
UCU	GCU	UAU	GAU	UCG	GCG	UAG	GAG
CUU	AUU	CGU	AGU	CUG	AUG	CGG	AGG
UUU	GUU	UGU	GGU	UUG	GUG	UGG	GGG

CCC	ACC	CCA	ACA	CAC	AAC	CAA	AAA
UCC	GCC	UCA	GCA	UAC	GAC	UAA	GAA
CCU	ACU	CCG	ACG	CAU	AAU	CAG	AAG
UCU	GCU	UCG	GCG	UAU	GAU	UAG	GAG
CUC	AUC	CUA	AUA	CGC	AGC	CGA	AGA
UUC	GUC	UUA	GUA	UGC	GGC	UGA	GGA
CUU	AUU	CUG	AUG	CGU	AGU	CGG	AGG
UUU	GUU	UUG	GUG	UGU	GGU	UGG	GGG

CCC	CAC	CCA	CAA	ACC	AAC	ACA	AAA
CUC	CGC	CUA	CGA	AUC	AGC	AUA	AGA
CCU	CAU	CCG	CAG	ACU	AAU	ACG	AAG
CUU	CGU	CUG	CGG	AUU	AGU	AUG	AGG
UCC	UAC	UCA	UAA	GCC	GAC	GCA	GAA
UUC	UGC	AUU	AGU	CUG	CGG	GUA	GGA
UCU	UAU	UCG	UAG	GCU	GAU	GCG	GAG
UUU	UGU	UUG	UGG	GUU	GGU	GUG	GGG

These six genomatrices have the following **phenomenological** properties:



CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

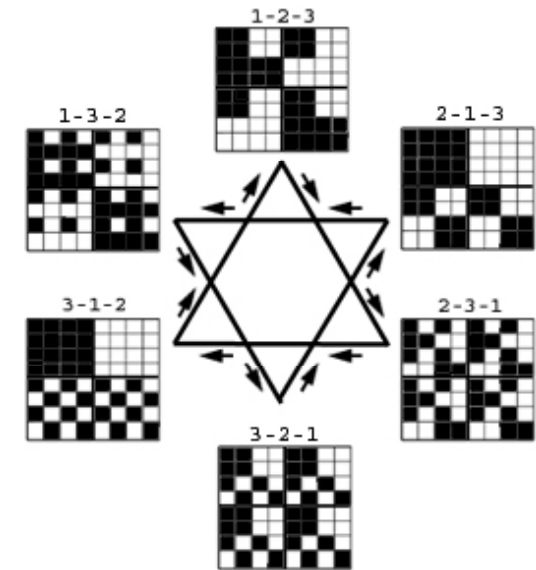
CCC	CAC	ACC	AAC	CCA	CAA	ACA	AAA
CUC	CGC	AUC	AGC	CUA	CGA	AUA	AGA
UCC	UAC	GCC	GAC	UCA	UAA	GCA	GAA
UUC	UGC	GUC	GGC	UUA	UGA	GUA	GGA
CCU	CAU	ACU	AAU	CCG	CAG	ACG	AAG
CCU	CGU	AUU	AGU	CUG	CGG	AUG	AGG
UCU	UAU	GCU	GAU	UCG	UAG	GCG	GAG
UUU	UGU	GUU	GGU	UUG	UGG	GUG	GGG

CC	CCA	ACC	ACA	CAC	CAA	AAC	AAA
CCU	CCG	ACU	ACG	CAU	CAG	AAU	AAG
UCC	UCA	GCC	GCA	UAC	UAA	GAC	GAA
UCU	UCG	GCU	GCG	UAU	UAG	GAU	GAG
CUC	CUA	AUC	AUA	CGC	CGA	AGC	AGA
CUU	CUG	AUU	AUG	CGU	CGG	AGU	AGC
UUC	UUA	GUC	GUA	UGC	UGA	GGC	GGA
UUU	UUG	GUU	GUG	UGU	UGG	GGU	GGG

CCC	ACC	CAC	AAC	CCA	ACA	CAA	AAA
UCC	GCC	UAC	GAC	UCA	GCA	UAA	GAA
CUC	AUC	CGC	AGC	CUA	AUA	CGA	AGA
UUC	GUC	UGC	GGC	UUA	GUA	UGA	GGA
CCU	ACU	CAU	AAU	CCG	ACG	CAG	AAG
UCU	GCU	UAU	GAU	UCG	GCG	UAG	GAG
CUU	AUU	CGU	AGU	CUG	AUG	CGG	AGG
UUU	GUU	UGU	GGU	UUG	GUG	UGG	GGG

CCC	ACC	CCA	ACA	CAC	AAC	CAA	AAA
UCC	GCC	UCA	GCA	UAC	GAC	UAA	GAA
CCU	ACU	CCG	ACG	CAU	AAU	CAG	AAG
UCU	GCU	UCG	GCG	UAU	GAU	UAG	GAG
CUC	AUC	CUA	AUA	CGC	AGC	CGA	AGA
CUU	GUC	UUA	GUA	UGC	GGC	UGA	GGA
CCU	AUU	CUG	AUG	CGU	AGU	CGG	AGG
UUU	GUU	UUG	GUG	UGU	GGU	UGG	GGG

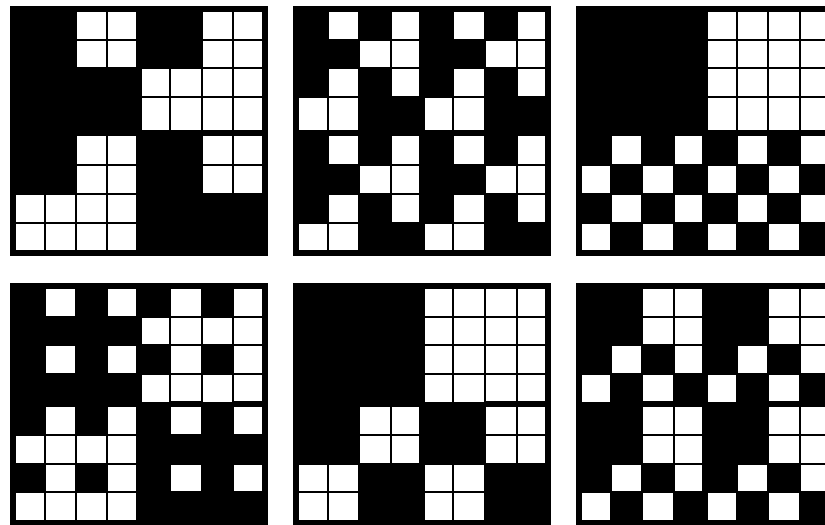
CCC	CAC	CCA	CAA	ACC	AAC	ACA	AAA
CUC	CGC	CUA	CGA	AUC	AGC	AUA	AGA
CCU	CAU	CCG	CAG	ACU	AAU	ACG	AAG
CUU	CGU	CUG	CGG	AUU	AGU	AUG	AGG
UCC	UAC	UCA	UAA	GCC	GAC	GCA	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GUA	GGA
UCU	UAU	UCG	UAG	GCU	GAU	GCG	GAG
UUU	UGU	UUG	UGG	GUU	GGU	GUG	GGG



1) every row of each of these mosaic matrices corresponds to one of Rademacher functions. It means that the connection of the genetic code with Rademacher functions is invariant relative to such **positional** permutations inside triplets. (This connection is invariant also relative to many cyclic **alphabetical** permutations  $C \rightarrow A \rightarrow G \rightarrow U \rightarrow C$ , etc.)



2) if we interpret each mosaic row as a relevant Rademacher function with its components “+1” and “-1”, six numeric (8\*8)-matrices arise:



where each black cell corresponds to “+1” and each white cell – to “-1”.



Unexpectedly these genomatrices present one of the main tools of theory of vector spaces – “**projective operators**”. By definition, a linear operator  $\mathbf{Y}$  in a linear space is named a projector operator if it satisfies the following condition:  $\mathbf{Y}^2 = \mathbf{Y}$ . But every of this six different matrices (multiplied by a general factor  $1/4$ ) satisfies this condition:

$$\mathbf{Y}_{123}^2 = \mathbf{Y}_{123} ,$$

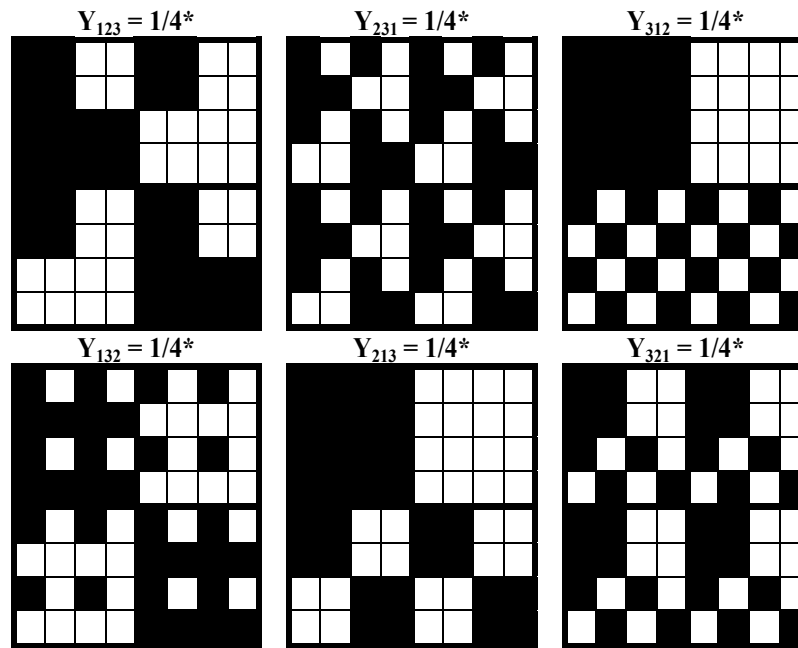
$$\mathbf{Y}_{231}^2 = \mathbf{Y}_{231} ,$$

$$\mathbf{Y}_{123}^2 = \mathbf{Y}_{123} ,$$

$$\mathbf{Y}_{123}^2 = \mathbf{Y}_{123} ,$$

$$\mathbf{Y}_{123}^2 = \mathbf{Y}_{123} ,$$

$$\mathbf{Y}_{123}^2 = \mathbf{Y}_{123} .$$





Theory of projective operators, which is used in quantum mechanics, automatic control systems, etc., is applied now in matrix genetics, which studies matrix presentations of genetic code systems. For example, some of these genetic projective operators are **commutative** ( $Y_{123} * Y_{321} = Y_{321} * Y_{123}$ , etc):

$$\begin{array}{l}
 Y_{312} * Y_{213} = Y_{213} * Y_{312} = \\
 Y_{123} * Y_{321} = Y_{321} * Y_{123} = \\
 Y_{231} * Y_{132} = Y_{132} * Y_{231} =
 \end{array}$$

Here each black cell corresponds to “+1/4”, each white cell – to “-1/4”, each grey cell – to “0”.

$$\begin{array}{l}
Y_{312} * Y_{213} = Y_{213} * Y_{312} = \\
Y_{123} * Y_{321} = Y_{321} * Y_{123} = \\
Y_{231} * Y_{132} = Y_{132} * Y_{231} =
\end{array}$$

Projectors in the right column are orthogonal each to another: their product is zero. It is known that every family of commutative projectors generates a **Boolean algebra of projectors**. Boolean algebras are the basis in computer technology and in many other fields. Now Boolean algebras of genetic projectors appear in bioinformatics. Families of commutative and non-commutative genetic projectors are studied intensively now.




### 8.3 The Genetic Code and Hadamard Matrices

- It is known that Rademacher functions are connected closely with Walsh functions which form a complete system of orthogonal functions and which are used widely in digital technologies, spectral analysis, noise-immunity coding, etc.
- We will show now a phenomenological fact that the described set of genomatrices with “Rademacher” rows is transformed by a natural genetic algorithm into a set of Hadamard matrices which contain complete systems of Walsh functions



By definition, a Hadamard matrix is a  $(n \times n)$ -matrix  $\mathbf{H}$  with entries  $\pm 1$  which satisfies  $\mathbf{H}^* \mathbf{H}^T = n \cdot \mathbf{E}$ , where  $\mathbf{H}^T$  - transposed matrix,  $\mathbf{E}$  – identity matrix. Some examples of Hadamard matrices are:

$$H_2 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}; H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}; H_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$



$$H_2 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}; H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}; H_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Rows of Hadamard matrices are Walsh functions. Any two rows of a Hadamard matrix are orthogonal. Kronecker product of any two Hadamard matrices is a Hadamard matrix again. Many normalized Hadamard matrices are unitary operators, and serve, in particular, for a creation of quantum computers, which are based on Hadamard gates (the evolution of a closed quantum system is unitary).

## Realization of Hadamard matrices in genomatrices

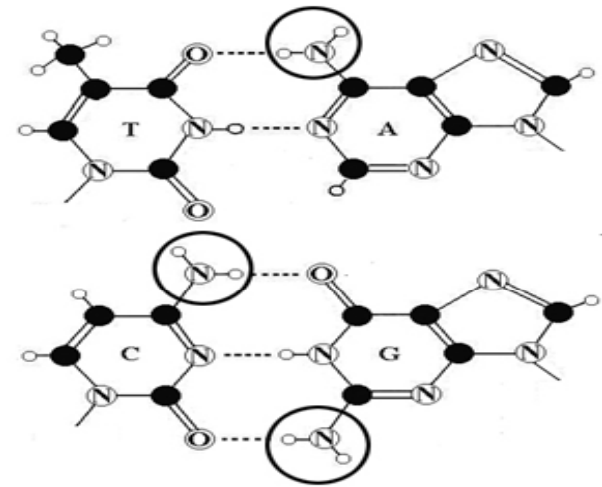
The genetic alphabet is characterized by the **ratio 3:1** because the letter U/T differs in the genetic alphabet from other three letters A, C, G phenomenologically:

- Firstly, at the transition from RNA to DNA only one letter U (uracil) is replaced by T (thymine), and the other three letters are not changed.

- Secondly, in the genetic alphabet, only the molecule U / T has not the important amino-group  $\text{NH}_2$ .

It resembles the **ratio 3:1** in the simplest Hadamard ( $2 \times 2$ )-matrices where one entry is opposite to other three entries by its sign:

+1	+1
-1	+1





This fact has led to a discovering of a hidden connection of genomatrices with Hadamard matrices by means of so called “U-algorithm” which transforms many described genetic matrices into Hadamard matrices. The “U-algorithm” is the following:

- each triplet in the black-and-white genomatrix of 64 triplets [C A; U G]<sup>(3)</sup> should change its own color into opposite color each time when the letter U/T stands in one of odd positions inside the triplet (in the first position or in the third position).

[C A; U G]<sup>(3)</sup>=

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
Pro	Pro	His	Gln	Thr	Thr	Asn	Lys
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
Pro	Pro	His	Gln	Thr	Thr	Asn	Lys
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
Leu	Leu	Arg	Arg	Ile	Met	Ser	Stop
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
Leu	Leu	Arg	Arg	Ile	Met	Ser	Stop
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
Ser	Ser	Tyr	Stop	Ala	Ala	Asp	Glu
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
Ser	Ser	Tyr	Stop	Ala	Ala	Asp	Glu
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
Phe	Leu	Cys	Trp	Val	Val	Gly	Gly
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG
Phe	Leu	Cys	Trp	Val	Val	Gly	Gly



Black	White	Black	White	Black	White	Black	White
White	Black	White	Black	White	Black	White	Black
Black	White	Black	White	Black	White	Black	White
White	Black	White	Black	White	Black	White	Black
Black	White	Black	White	Black	White	Black	White
White	Black	White	Black	White	Black	White	Black
Black	White	Black	White	Black	White	Black	White
White	Black	White	Black	White	Black	White	Black





## HADAMARD MATRICES

- Hadamard matrices are one of effective tools of noise-immunity coding in digital communication (the error-correcting code by Reed-Muller, Hadamard codes, etc.). They are used in quantum computers with Hadamard gates, in quantum mechanics as unitary operators, in multi-channels spectrometers with Hadamard transformations, etc.
- The phenomenological facts of connections of genetic structures with special classes of projective operators, Hadamard matrices, etc. testify that the **genetic code is algebraic in its nature**. The results of matrix genetics allow developing new methods in bioinformatics and in applied fields of genetic algorithms.



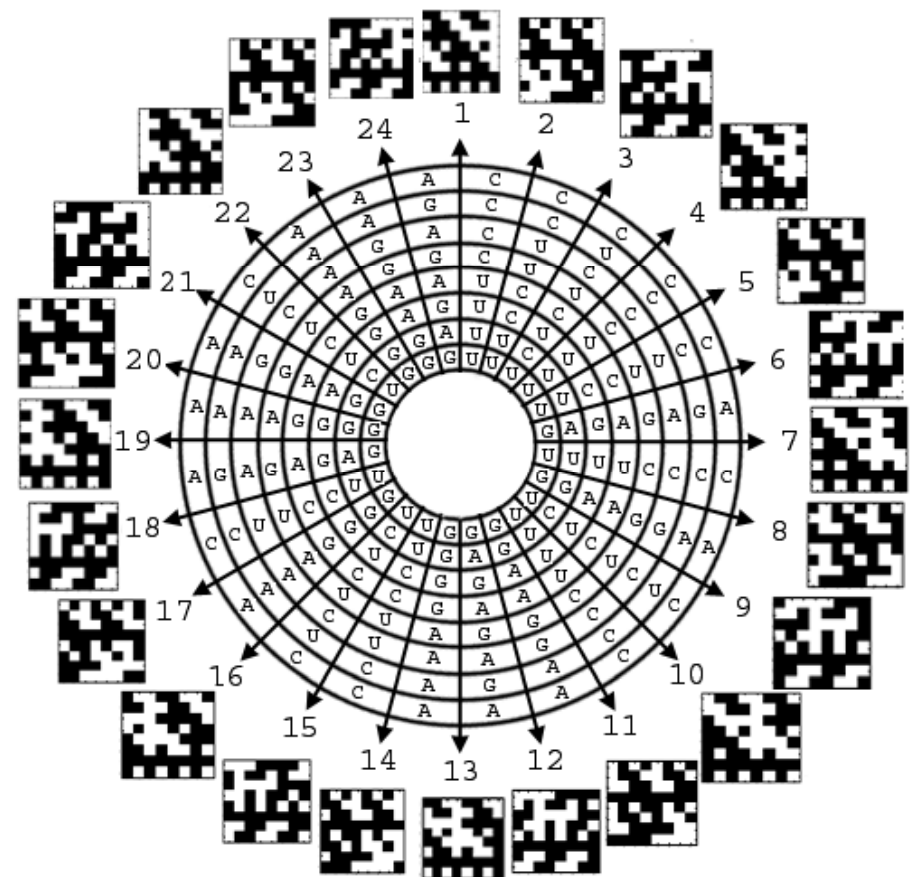
## HADAMARD MATRICES

- Discrete character of the genetic code shows that genetic informatics is based on principles of combinatorics. Our results of matrix genetics clarify that genetic combinatorics is a combinatorics of finite-dimensional discrete vector spaces.

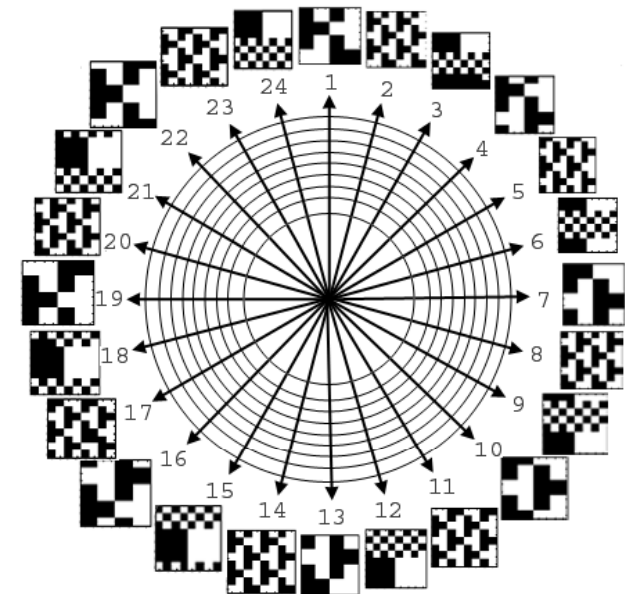
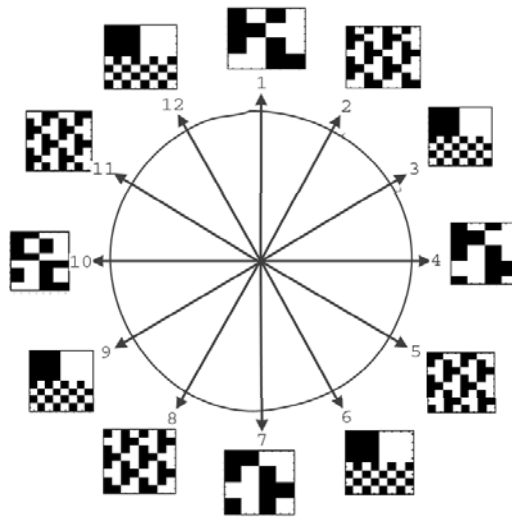
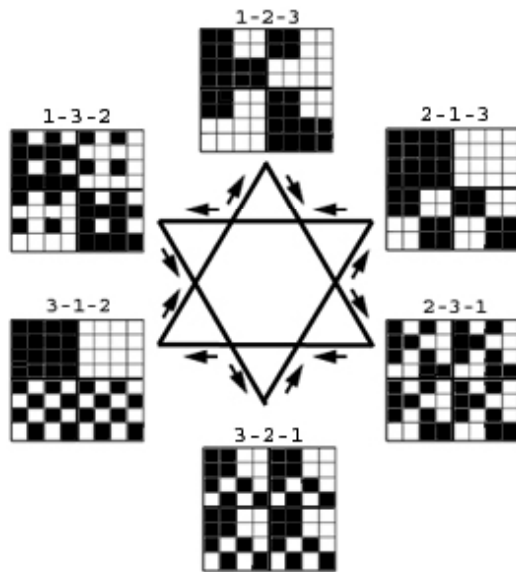
The Mendel's law of independent inheritance of traits demonstrates that molecular-genetic information defines traits on macro-levels of biological bodies. It testifies that biological organisms are some combinatorial algorithmic machines. Matrix genetics aims to reveal secrets of these algorithmic biomachines.

Matrix genetics has discovered that many families of numeric genetic matrices present hierarchies of cyclic groups of transformations. For example each described Hadamard genomatrix  $\mathbf{H}$  forms a cyclic group  $\mathbf{H}^n$  with a period 24:  
 $\mathbf{H}^n = \mathbf{N}^{n+24}$ .

Another example:  
 cyclic permutations of the genetic letters along rows of matrices of 64 triplets [C A; U G]<sup>(3)</sup> lead to a cyclic group of 24 Hadamard genomatrices:



So called “genetic dials (clock faces)” and new approaches for modelling of cyclic processes arise in matrix genetics. They concern the problem of biological time and of inherited biological clock additionally.



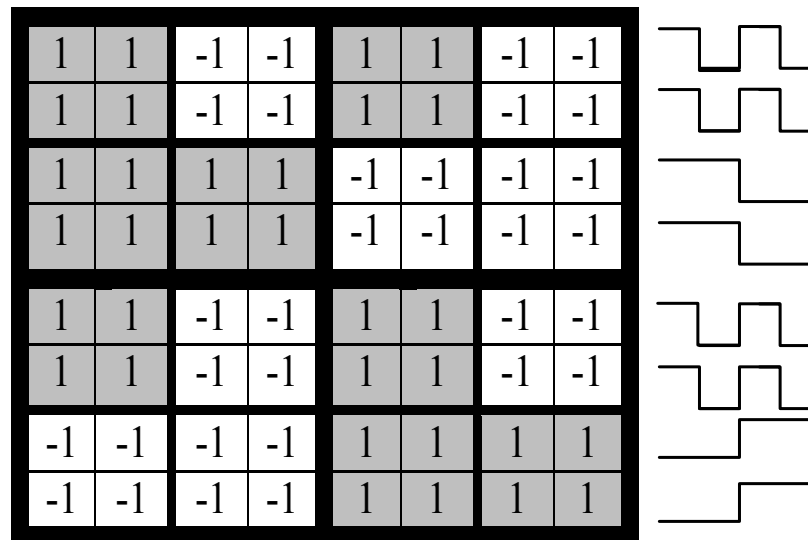


## 8.4 GENETIC YIN-YANG ALGEBRAS

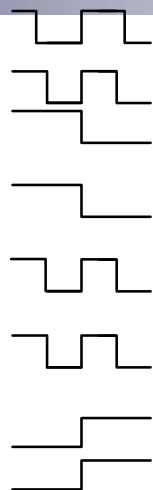
- A few words about new **genetic Yin-Yang algebras** which simulate some properties of the genetic code and its biological evolution. A wide set of these Yin-Yang algebras, which form cyclic groups, was revealed in matrix genetics in a course of study of phenomenological properties of the genetic code.
- Let us begin with the described Rademacher form  $Y_8$  of the initial genomatrix  $[C \ A; U \ G]^{(3)}$ :



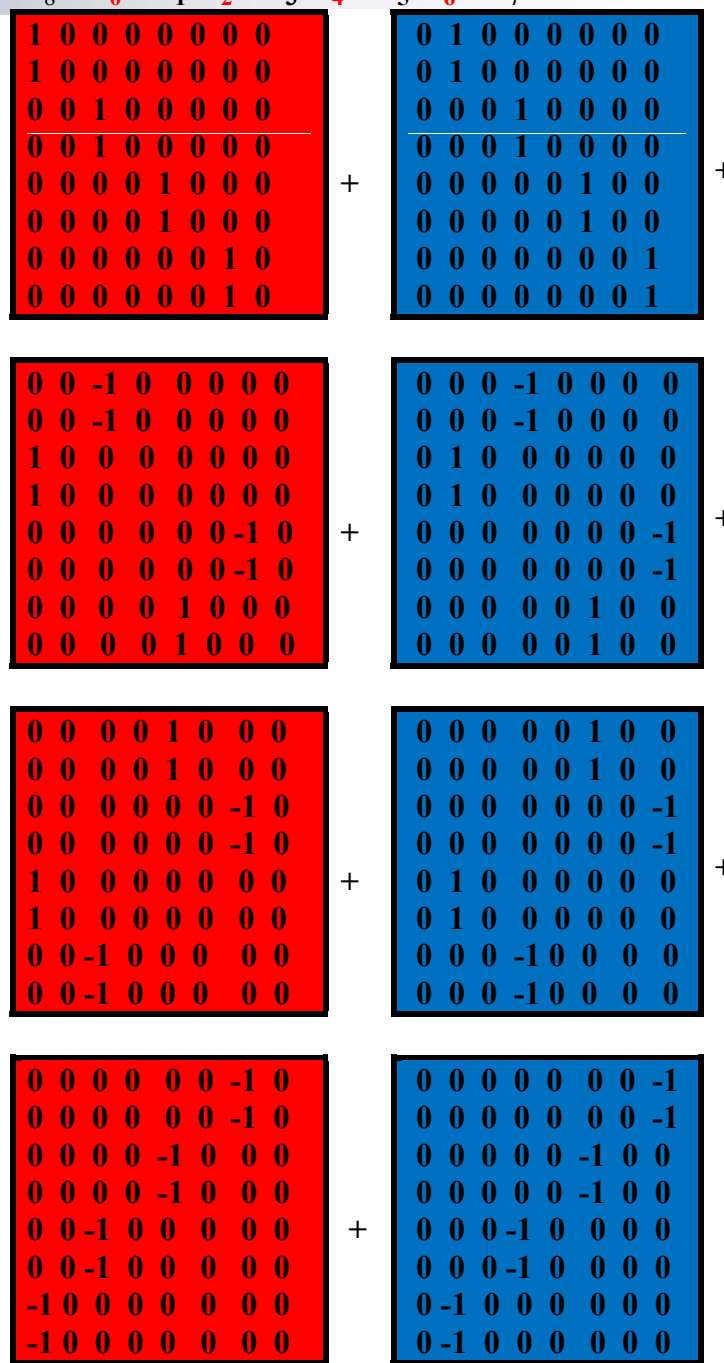
Let us begin with the described Rademacher form  $Y_8$  of the initial genomatrix  $[C \ A; U \ G]^{(3)}$ :



+1	+1	-1	-1	+1	+1	-1	-1
+1	+1	-1	-1	+1	+1	-1	-1
+1	+1	+1	+1	-1	-1	-1	-1
+1	+1	+1	+1	-1	-1	-1	-1
+1	+1	-1	-1	+1	+1	-1	-1
+1	+1	-1	-1	+1	+1	-1	-1
-1	-1	-1	-1	+1	+1	+1	+1
-1	-1	-1	-1	+1	+1	+1	+1



$$Y_8 = f_0 + m_1 + f_2 + m_3 + f_4 + m_5 + f_6 + m_7 =$$



This matrix  $Y_8$  is a sum of 8 sparse matrices  $f_0, m_1, f_2, m_3, f_4, m_5, f_6, m_7$ , the set of which is closed relative multiplication. It means that product of any two matrices from this set gives a matrix which belong to this set again. (Here pink color marks matrices with even indexes 0, 2, 4, 6, and blue color marks matrices with odd indexes 1, 3, 5, 7).

The table of multiplication of these 8 matrices

$f_0, m_1, f_2, m_3, f_4, m_5, f_6, m_7$  is the following:

$$Y_8 = f_0 + m_1 + f_2 + m_3 + f_4 + m_5 + f_6 + m_7 =$$

1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0

0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0

0	0	-1	0	0	0	0	0	0
0	0	-1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	-1	0	0
0	0	0	0	0	0	-1	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0

0	0	0	-1	0	0	0	0	0
0	0	0	-1	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	-1	0
0	0	0	0	0	0	0	-1	0
0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0

0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	-1	0	0
0	0	0	0	0	0	-1	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0	0
0	0	-1	0	0	0	0	0	0

0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	-1	0
0	0	0	0	0	0	0	-1	0
0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	0	-1	0	0	0	0	0
0	0	0	-1	0	0	0	0	0

0	0	0	0	0	0	0	-1	0
0	0	0	0	0	0	0	-1	0
0	0	0	0	-1	0	0	0	0
0	0	0	0	-1	0	0	0	0
0	0	-1	0	0	0	0	0	0
0	0	-1	0	0	0	0	0	0
-1	0	0	0	0	0	0	0	0
-1	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0	-1
0	0	0	0	0	0	-1	0	0
0	0	0	0	0	0	-1	0	0
0	0	0	-1	0	0	0	0	0
0	0	0	-1	0	0	0	0	0
0	0	-1	0	0	0	0	0	0
0	0	-1	0	0	0	0	0	0

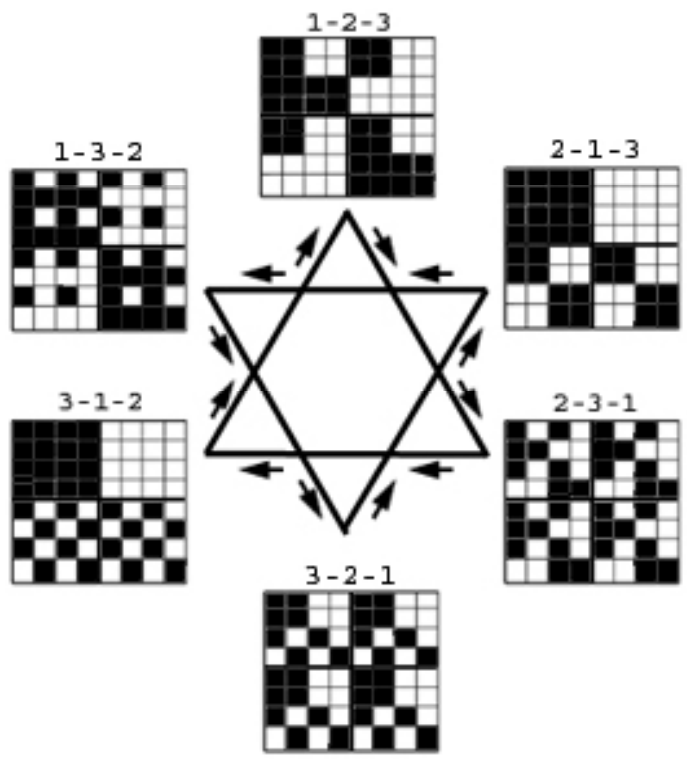
	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_0$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$m_1$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_2$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_6$	$-m_7$	$f_4$	$m_5$
$m_3$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_6$	$-m_7$	$f_4$	$m_5$
$f_4$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$m_5$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$f_6$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_2$	$-m_3$	$f_0$	$m_1$
$m_7$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_2$	$-m_3$	$f_0$	$m_1$

	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_0$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$m_1$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_2$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_6$	$-m_7$	$f_4$	$m_5$
$m_3$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_6$	$-m_7$	$f_4$	$m_5$
$f_4$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$m_5$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$f_6$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_2$	$-m_3$	$f_0$	$m_1$
$m_7$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_2$	$-m_3$	$f_0$	$m_1$

It means that one can say about the 8-dimensional algebra of matrix operators  $f_0, m_1, f_2, m_3, f_4, m_5, f_6, m_7$ . This table of multiplication shows that in this algebra the set of all basic matrices and coordinates are divided into two subsets – with even indexes  $0, 2, 4, 6$  (**Yin subset**) and with odd indexes  $1, 3, 5, 7$  (**Yang subset**). In this reason we name such algebra as **Yin-Yang algebra** (or even-odd algebra, or bipolar algebra).



Every of  $(8 \times 8)$ -genomatrices, which are formed from the initial matrix  $[C A; G U]^{(3)}$  by means of all six possible positional permutations in triplets, corresponds to its own Yin-Yang algebra with its own multiplication table:



	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_0$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$m_1$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_2$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$m_3$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$f_4$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$m_5$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$f_6$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$
$m_7$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$

	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_6$	$m_6$	$m_7$
$f_0$	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_6$	$m_6$	$m_7$
$f_1$	$f_1$	$-f_0$	$-f_3$	$f_2$	$m_6$	$-m_4$	$-m_7$	$m_6$
$f_2$	$f_2$	$f_3$	$f_0$	$f_1$	$m_6$	$m_7$	$m_4$	$m_5$
$f_3$	$f_3$	$-f_2$	$-f_1$	$f_0$	$m_7$	$-m_6$	$-m_6$	$m_4$
$m_4$	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_6$	$m_6$	$m_7$
$m_6$	$f_1$	$-f_0$	$-f_3$	$f_2$	$m_6$	$-m_4$	$-m_7$	$m_6$
$m_6$	$f_2$	$f_3$	$f_0$	$f_1$	$m_6$	$m_7$	$m_4$	$m_5$
$m_7$	$f_3$	$-f_2$	$-f_1$	$f_0$	$m_7$	$-m_6$	$-m_6$	$m_4$

	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_0$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$m_1$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_2$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$m_3$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$f_4$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$m_5$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$f_6$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$
$m_7$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$

	$f_0$	$f_1$	$m_2$	$m_3$	$f_4$	$f_5$	$m_6$	$m_7$
$f_0$	$f_0$	$f_1$	$m_2$	$m_3$	$f_4$	$f_5$	$m_6$	$m_7$
$f_1$	$f_1$	$-f_0$	$m_3$	$-m_2$	$-f_5$	$f_4$	$-m_7$	$m_6$
$m_2$	$f_0$	$f_1$	$m_2$	$m_3$	$f_4$	$f_5$	$m_6$	$m_7$
$m_3$	$f_1$	$-f_0$	$m_3$	$-m_2$	$-f_5$	$f_4$	$-m_7$	$m_6$
$f_4$	$f_4$	$f_5$	$m_6$	$m_7$	$f_0$	$f_1$	$m_2$	$m_3$
$f_5$	$f_4$	$f_5$	$m_6$	$m_7$	$f_0$	$f_1$	$m_2$	$m_3$
$m_6$	$f_4$	$f_5$	$m_6$	$m_7$	$f_0$	$f_1$	$m_2$	$m_3$
$m_7$	$f_5$	$-f_4$	$m_7$	$-m_6$	$-f_1$	$f_0$	$-m_3$	$m_2$

	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_0$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$m_1$	$f_0$	$m_1$	$f_2$	$m_3$	$f_4$	$m_5$	$f_6$	$m_7$
$f_2$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$m_3$	$f_2$	$m_3$	$-f_0$	$-m_1$	$-f_4$	$-m_5$	$f_6$	$m_7$
$f_4$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$m_5$	$f_4$	$m_5$	$f_6$	$m_7$	$f_0$	$m_1$	$f_2$	$m_3$
$f_6$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$
$m_7$	$f_6$	$m_7$	$-f_4$	$-m_5$	$-f_0$	$-m_1$	$f_2$	$m_3$

	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_5$	$m_6$	$m_7$
$f_0$	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_5$	$m_6$	$m_7$
$f_1$	$f_1$	$-f_0$	$-f_3$	$f_2$	$m_5$	$-m_4$	$-m_7$	$m_6$
$f_2$	$f_2$	$-f_3$	$-f_0$	$f_1$	$m_5$	$-m_7$	$-m_4$	$m_6$
$f_3$	$f_2$	$-f_3$	$-f_0$	$f_1$	$m_5$	$-m_7$	$-m_4$	$m_6$
$m_4$	$f_0$	$f_1$	$f_2$	$f_3$	$m_4$	$m_5$	$m_6$	$m_7$
$m_5$	$f_1$	$-f_0$	$-f_3$	$f_2$	$m_5$	$-m_4$	$-m_7$	$m_6$
$m_6$	$f_2$	$-f_3$	$-f_0$	$f_1$	$m_5$	$-m_7$	$-m_4$	$m_6$
$m_7$	$f_3$	$-f_2$	$-f_1$	$f_0$	$m_7$	$-m_6$	$-m_5$	$m_4$



## 8.5 CHALLENGES AND PERSPECTIVES

- The discovery of connections of the genetic matrices with Hadamard matrices leads to many new possible investigations using the methods of symmetry, of spectral analysis, etc.
- It seems that investigations of structural and functional principles of bio-information systems from the viewpoint of quantum computers and of unitary Hadamard operators are very promising.
- A comparison of orthogonal systems of Walsh functions in molecular-genetic structures and in genetically inherited macro-physiological systems can give new understanding to the interrelation of various levels in biological organisms.
- Data about the genetic Hadamard matrices together with data about algebras of the genetic code can lead to new understanding of genetic code systems, to new effective algorithms of information processing and, perhaps, to new directions in the field of quantum computers.



## **Part II Biological Functions, Networks, Systems Biology and Cognitive Informatics**

---

---

### **9. Bioinformatics, Denotational Mathematics, and Cognitive Informatics**

- ❖ Introduction
- ❖ Emerging Pattern, Dissipative Structure, and Evolving Cognition
- ❖ Denotational Mathematics and Cognitive Computing
- ❖ Challenges and Perspectives



## 9.1 INTRODUCTION

- Bioinformatics is the comprehensive applications of mathematics, science, and a core set of problem-solving methods to the understanding of living systems. It will have profound impacts on all fields of biological and medical sciences. Cognition is viewed as a process of living systems.
- Cognition is an abstract property of advanced living organisms. It is studied as a direct property of a brain or of an abstract mind on sub-symbolic and symbolic levels.



## 9.1 Introduction

- Cognitive informatics studies cognition and information sciences that investigates the processes of the natural intelligence. As both fields continue their rapid development and progress, it is a central challenge to understand the biological basis of cognition, perception, learning, memory, thought, and mind.
- Patterns, structures, and rules arise and play an important role in living systems and nearly all branches of science. Living systems are open self-organizing systems that have the special characteristics of life and interact with their environment. This takes place by means of information and material-energy exchanges.



## 9.2 EMERGING PATTERN, DISSIPATIVE STRUCTURE, AND EVOLVING COGNITION

- **Autopoiesis-The Pattern of Life:** Autopoiesis literally means "auto (self)-creation" and expresses a fundamental interaction between structure and function. Autopoiesis is a network pattern in which the function of each component involves with the production or transformation of other components in the network. The simplest living system we know is the biological cell. The eukaryotic cell, for example, is made of various biochemical components such as nucleic acids and proteins, and is organized into bounded structures such as the cell nucleus, various organelles, a cell membrane and cytoskeleton.



## 9.2 EMERGING PATTERN, DISSIPATIVE STRUCTURE, AND EVOLVING COGNITION

- **Dissipative Structure-the Structure of Living Systems:** The term dissipative structure of a living system was coined by Ilya Prigogine who pioneered research in the field of thermodynamics in (Mingers, 1994). A dissipative structure is a thermodynamically open system to the flow of energy and matter. A dissipative structure is operating far from thermodynamic equilibrium in an environment with which it exchanges energy and matter.
- **Cognition-the Process of Life:** The concept of cognition is closely related to such abstract concepts as mind, reasoning, perception, intelligence, learning, and many others that describe numerous capabilities of the human mind and expected properties of artificial or synthetic intelligence.



### 9.3. DENOTATIONAL MATHEMATICS AND COGNITIVE COMPUTING

- Denotational mathematics is a category of expressive mathematical structures that deals with high level mathematical entities beyond numbers and sets, such as abstract objects, complex relations, behavioral information, concepts, knowledge, processes, and systems.
- Cognitive informatics studies intelligent behavior and cognition. Cognition includes mental states and processes, such as thinking, reasoning, learning, perception, emotion, consciousness, remembering, language understanding and generation, etc. In the emerging theory of living systems mind is not a thing, but a process. It is cognition, the process of knowing, and it's identified with the process of life itself.



### 9.3. DENOTATIONAL MATHEMATICS AND COGNITIVE COMPUTING

<b>Basic power expressive in system modeling</b>	<b>Classic mathematics</b>	<b>Denotational mathematics</b>	<b>Usage</b>
<i>To be</i>	Logic	Concept Algebra	Identify objects and attributes
<i>To have</i>	Set theory	System Algebra	Describe relations and procession
<i>To do</i>	Functions	Real-Time Process Algebra	Describe status and behavior

## COGNITIVE INFORMATICS

Cognitive Computing	Computational Intelligence	Neural Informatics
Informatics models of the brain	Imperative vs. autonomous Computing	Neuroscience foundations of information processing
Cognitive processes of the brain	Reasoning and inferences	Cognitive models of the brain
Internal information processing Mechanisms	Cognitive informatics foundations	Functional modes of the brain
Theories of natural intelligence	Robotics	Neural models of memory
Intelligent foundations of computing	Informatics foundations of software engineering	Neural networks
Denotational mathematics	Fuzzy/rough sets/logic	Neural computation
Abstraction and means	Knowledge engineering	Cognitive linguistics
Ergonomics	Pattern and signal recognitions	Neuropsychology
Informatics laws of software	Autonomic agent technologies	Bioinformatics

## COGNITIVE INFORMATICS

Cognitive Computing	Computational Intelligence	Neural Informatics
Knowledge representation	Memory models	Biosignal processing
Models of knowledge and skills	Software agent systems	Cognitive signal processing
Formal linguistics	Decision theories	Gene analysis and expression
Cognitive complexity & metrics	Problem solving theories	Cognitive metrics
Distributed intelligence	Machine learning systems	Neural signal interpretation
Semantic computing	Distributed objects/granules	Visual information representation
Emotions/motivations/attitudes	Web contents cognition	Visual semantics
Perception and consciousness	Nature of software	Sensational cognitive processes
Hybrid (AI/NI) intelligence	Granular computing	Human factors in systems



## 9.4 CHALLENGES AND PERSPECTIVES

- From a scientific perspective discovering how the brain thinks is a major undertaking in the history of mankind.
- Bioinformatics provides computational and experimental tools to study the biological patterns, structures, and functions.
- Cognitive informatics investigates the internal information processing mechanisms and process of life-cognition.



## 9.4 CHALLENGES AND PERSPECTIVES

- “Understanding the human mind in biological terms has emerged as the central challenge for science in the twenty first century. We want to understand the biological nature of perception, learning, memory, thought, consciousness, and the limits of free will,” as Kandel put it in (Kandel, 2006) “Thus, we gain from the new science of mind not only insights into ourselves-how we perceive, learn, remember, feel, and act-but also a new perspective of ourselves in the context of biological evolution.”
- “The task of neural science is to explain behavior in terms of the activities of the brain. How does the brain marshal its millions of individual nerve cells to produce behavior, and how are these cells influenced by the environment...? The last frontier of the biological sciences – their ultimate challenge – is to understand the biological basis of consciousness and the mental processes by which we perceive, act, learn, and remember.”  
(Kandel)



## **Part II Biological Functions, Networks, Systems Biology and Cognitive Informatics**

---

---

### **10. The Evolutionary Trends and Central Dogma of Informatics**

- ❖ Introduction
- ❖ Evolutionary Trends of Informatics
- ❖ Central Dogma of Informatics
- ❖ Challenges and Perspectives



## 10.1 INTRODUCTION

- Life consists of matter and energy, but it is not just matter and energy. Life is also information.
- Life has three fundamental dimensions.
- The life of an individual comes from the DNA of its parents. DNA is insignificant in terms of its elemental composition. It's composed of nitrogen, oxygen, sulfur, etc. DNA as a source of energy is composed of the similar level of chemical energy that can be produced by experiments. The characteristic of DNA is an informational molecule, a molecule containing a large amount of information.
- Informatics is evolving and being transformed. Many boundaries among science, engineering and social systems are cross-linked in the face of combinations of knowledge and tools.



## 10.1 INTRODUCTION

- Informatics studies the foundation, representation, processing, and communication of information in natural and artificial systems. The central notion is the transformation of data to information and information to knowledge - whether by computation or communication, whether by organisms or artifacts. It deals with the structure, function, behavior, and interactions of natural and artificial computational systems. It has computational, experimental, theoretical, cognitive and social aspects.
- Three of the truly fundamental questions of science are: "What is matter?", "What is life?" and "What is mind?" The physical and biological sciences concern the first two. The emerging science of Informatics contributes to our understanding of the latter two by providing a basis for the study of organization and process in biological and cognitive systems. Progress can best be made by means of strong links with the existing disciplines devoted to particular aspects of these questions.



## 10.2 EVOLUTIONARY TRENDS OF INFORMATION SCIENCES

Physical information science and technology	Non-organic/nonliving material-based
Biological information science and technology	Organic/living material-based
Societal information science and technology	Language/mind-based



## 10.2 EVOLUTIONARY TRENDS OF INFORMATION SCIENCES

**Physical information science and technology** is viewed as an original system science that exists in the physical world. They are made of physical materials with informational instructions. A voluminous literature on physical science and technology is documented throughout history (Meyers, 2001).

**Biological information science and technology** is defined as biological means embedded in living systems. It is made of biological matter and naturally evolved “machines” that perform molecular calculations and complex functions. The biological or natural technology is a major and most effective technology ensuring life on our planet. And acquirement of this technology, occurring in modern time, is a major movement in the evolution of mankind.



## 10.2 EVOLUTIONARY TRENDS OF INFORMATION SCIENCES

**Societal information science and technology**, the technological or engineering counterpart to the social sciences, emphasizes "genuine solutions" to social problems and social life, treating their underlying causes. It is a science that answers questions about "what" and "why"; since technology is the collection of study, invention and refinement of tools and techniques, and therefore it answers "how" questions. Law, government, business, finance, research, development, education, and other activities within our human society is a collection of tools/rules and techniques applied for societal purposes, and thus societal technology. These three phases of technologies are well connected with the process of universal evolution. Evolution is the primary cosmic force which creates the order and makes it visible in different categories of nature.

## UNIVERSAL EVOLUTION

<b>PHYSICAL SPHERE</b>	<b>BIOSPHERE</b>	<b>METABIOSPHERE</b>
Physical matter  Elementary particles  -Atoms -Molecules ---  <i>Pre-biological Evolution</i>	Living matter  Replicating molecules  -Cells -Organisms ---  <i>Biological Evolution</i>	Human matter  Human mind  -Human Culture -Human Morality ---  <i>Meta-biological Evolution</i>



## CONVERGENCE OF SCIENCES: INFORMATION

- Many boundaries among science, engineering and social systems are cross-linked in the face of combinations of knowledge and tools.
- The **cells** contain millions of molecular machines connected by biochemical reactions.
- The **brain** is a natural network technology of nerve cells connected by axons.
- **Societies** are social network technology of people linked by various relations such as familial trees, friendships, mutual ties, and social networks.
- The **natural language** we are using to express our thoughts is a communication technology that connecting words by syntactic relationships.



## CONVERGENCE OF SCIENCES: INFORMATION

### The Emergence of Holistic Thinking

- **Complexes** of interrelated processes on multiple levels/a general air of being insoluble (Multidimensional databases). In the world of quantum physics everything is **interconnected**.
- The **convergence of CS and biology** (cellular automata, neural networks, genetic algorithms, algorithmic systems biology, etc., ) will serve both disciplines, providing each with greater power and relevance.
- The **convergence of social and technological networks**: Facebook, LinkedIn, MySpace, Wikipedia, YouTube...
- The interactions among these physical, biological, and societal information sciences and technologies are connected and interdependent. These technologies are the driving force of life. They are connected through a central unit: *Information*.
- **System = Things • Relations (S = T•R)**



## 10.3 CENTRAL DOGMA OF INFORMATICS

- The central dogma of molecular biology was first enunciated by Francis Crick in 1958 and re-stated in a Nature paper published in 1970 (Crick, 1970). It states the conversion of the genetic message in DNA to a functional mRNA (transcription) and subsequent conversion of the copied genotype to a phenotype in the form of proteins. The process of conversion of a mRNA to a functional protein is known as translation.
- This central dogma has paved a revolutionary road for further investigations in great details of the process of replication, transcription, and translation. Furthermore, this central dogma has provided a great impact on the study of informatics. The process of conversion from data to information and to knowledge has a much broader and fundamental impact on many domains as the information is an intrinsic property of universe-intricate organization of matter and energy.



## WHAT IS INFORMATION?

- According to **Peirce**, information was embedded in his wider theory of **symbolic communication**. It integrates the aspects of signs and expressions.
- Claude E. **Shannon**, for his part, was very cautious: “The word ‘information’ has been given different meanings by various writers in the general field of information theory. It can be viewed as flows of mass-energy **forms**.”
- According to **Tom Stonier**, information is **an implicit component** of virtually every single equation governing the laws of nature. Information is the core ingredient of all communication and control processes in living and nonliving systems. It was propagated as:
  - Patterns of light (from the book to the eye).
  - Pulses of membrane depolarization (from the eye to the brain).
  - Pulses of chemical substances (between individual nerves).
  - Pulses of light in optical fibers.



## WHAT IS INFORMATION?

- According to **Floridi**, four kinds of mutually compatible phenomena are commonly referred to as "information":
  - Information **about something** (e.g. a train timetable)
  - Information **as something** (e.g. DNA, or fingerprints)
  - Information **for something** (e.g. algorithms or instructions)
  - Information **in something** (e.g. a pattern or a constraint).
- **Gregory Bateson** defined information as “a **difference** that makes a difference”



## 10.3 CENTRAL DOGMA OF INFORMATICS

- **Data:** Here we define the data as a collection of symbols, images, or other outputs from devices on source to represent the qualitative or quantitative attributes of a variable or set of variables that are unprocessed and no relationships established.
- **Information:** Here we define information as a multidimensional organized entity derived from data source through a process of transformation. The information entity has many attributes or patterns with relationships. It has its own structure and can be converted to knowledge through a process of transformation.



## 10.3 CENTRAL DOGMA OF INFORMATICS

- **Knowledge:** Here we define knowledge as a multidimensional functional entity derived from information through a process of transformation. Knowledge has its own structure and function. Knowledge about primitive organisms provides much information about shared metabolic features, and hints at diseases that affect humans in an economical and ethically acceptable manner. Knowledge from many scientific disciplines and their subfields has to be integrated to achieve the goals of informatics. Applying knowledge can lead to new scientific methods, to new diagnostics and to new discoveries.

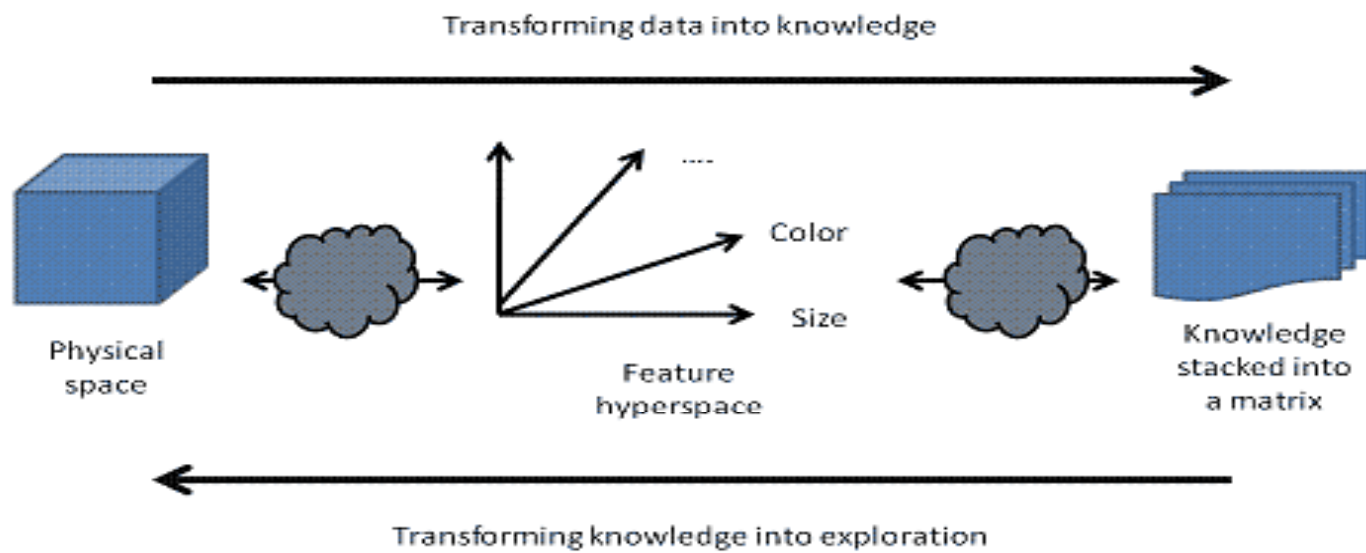


## CENTRAL DOGMA OF INFORMATICS

<b>Data</b>	<b>Information</b>	<b>Knowledge</b>
<p>A datum is a small chunk of information.</p>	<p>Information is thought of as organized data, or 'facts' organized into a coherent patterns.</p> <p><b>Information = Data + Meaning</b></p> <p>(Davenport and Prusak)</p>	<p>Knowledge is thought of as organized and internalized information and ability to utilize the information.</p> <p><b>Knowledge = Internalized information + Ability to utilize the information</b></p> <p>(Davenport and Prusak)</p>

# CENTRAL DOGMA OF INFORMATICS

Transforming Data into Information and into Knowledge





## CENTRAL DOGMA OF INFORMATICS

<b>Data (Replication)</b>	<b>Information (Transformation)</b>	<b>Knowledge (Application)</b>
-Data Structure -Data Mining -Data Analysis -Data Integration -Data Replication -Databases ---	-Pattern Discovery -Information Processing -Information Storage -Information Retrieval -Information Flow -Information Control ----	-Knowledge Discovery -Knowledge Management -Knowledge Transfer ---



## 10.4 CHALLENGES AND PERSPECTIVES

- What are the structures of data?
- What are the topological and geometrical properties of information? What are the functional features of knowledge?
- How are data transcribed into information? How are information sets translated into knowledge?



## 10.4 CHALLENGES AND PERSPECTIVES

- Data mining is an active field that examines the process of extracting hidden patterns from data.
- Knowledge discovery is a growing field to examine the process of converting the information into knowledge. Knowledge tends to travel from descriptive to qualitative to quantitative.
- During the process of this transition, numerical, graphical and mathematical aspects emerge to explore the issues of data, knowledge, intelligence, noise and meaning. This new kind of mathematics may be called the mathematics of knowledge and intelligence.



## TEN CHALLENGES IN THE SYNERGY BETWEEN BIOLOGY AND MATHEMATICS (1)

### **Five biological challenges that could stimulate, and benefit from, major innovations in mathematics**

1. Understand cells, their diversity within and between organisms, and their interactions with the biotic and abiotic environments.
2. Understand the brain, behavior, and emotion.
3. Replace the tree of life with a network or tapestry to represent lateral transfers of heritable features such as genes, genomes, and prions
4. Couple atmospheric, terrestrial, and aquatic biospheres with global physicochemical processes.
5. Monitor living systems to detect large deviations such as natural or induced epidemics or physiological or ecological pathologies.



## TEN CHALLENGES IN THE SYNERGY BETWEEN BIOLOGY AND MATHEMATICS (2)

### **Five mathematical challenges that would contribute to the progress of biology**

1. Understand computation. Find more effective ways to gain insight and prove theorems from numerical or symbolic computations and agent-based models.
2. Find better ways to model multi-level systems, for example, cells within organs within people in human communities in physical, chemical, and biotic ecologies.
3. Understand probability, risk, and uncertainty.
4. Understand data mining, simultaneous inference, and statistical de-identification.
5. Set standards for clarity, performance, publication and permanence of software and computational results.

# NEW PARADIGM OF COMPUTATIONAL SCIENCES...

<b>MATHEMATICAL THEORY OF COMPUTATION</b>	<ul style="list-style-type: none"> <li>➤ Information Algebra and Structure</li> <li>➤ Information Geometry and Architecture</li> <li>➤ Information Topology...</li> </ul>
<b>COMPUTATIONAL INFORMATION SCIENCE AND SYSTEMS</b>	<ul style="list-style-type: none"> <li>➤ Biological and Physical Information Systems</li> <li>➤ Social and behavioral Information Systems</li> <li>➤ Medical and health Information Systems</li> <li>➤ Geographical/Environmental Information System</li> <li>➤ Educational Information Systems</li> <li>➤ Business Management Information Systems...</li> </ul>
<b>COMPUTATION SYSTEMS ENGINEERING</b>	<p>Phases: Analysis/Design/Implementation/Operation</p> <ul style="list-style-type: none"> <li>➤ Integrative Information Systems Engineering</li> <li>➤ IBM Roadrunner and China Godson 3 Chip</li> </ul>
<b>COMPUTATION TECHNOLOGIES/INTELLIGENT MACHINE</b>	<ul style="list-style-type: none"> <li>➤ Computing Devices/Wireless Devices/</li> <li>➤ Embedded Devices/Remote-Sensor Devices/</li> <li>➤ Living Machines (Cell to Cell)</li> </ul>
<b>COMPUTATION ENGINEERING AND SOCIETY</b>	<ul style="list-style-type: none"> <li>➤ Impact and Implication on Society/</li> <li>➤ Legal and Ethical Issues/</li> <li>➤ Cultural and Social Impact...</li> </ul>

## GENETIC MUSIC (GENE, CULTURE, AND MIND)

<http://www.toshima.ne.jp/~edogiku/>





**THANK YOU!**

$\iiint \{M(\text{xyz})+C(0\&1)+B(\text{dna})\} \partial(\text{info})$