Information and Life

Dr Giuditta Franco

Department of Computer Science, University of Verona, Italy
giuditta.franco@univr.it

October 18, 2017

Fundamentals

Formalisms: contexts interpretative to give meaning to observations. Example: 00101101.

Basic concepts of set, function/relation, iteration (local dynamics), sequence and multiset (as functions). Polinomial representation (balanced chemical reactions) and support. (Multi)Set cardinality, string length, and absolute value.

Functions: arity, injective, surjective, bijective (one-to-one), inverse function, partial, total, monotone, composition of functions. How many are boolean functions on A (2^n if n = |A|)? (In general, are m^n, n=|domain| and m=|image|). How many, in general, the one-to-one functions?

Information and Life

Information are data (on a physical support) which are *stored, transformed, transmitted*. Data, recognized by agents, guide their action on data. Information is discrete when it is represented on discrete mathematical structures.

Information is *data transformation* just as energy is *state transformation* - computer is digitalization of mathematics, as DNA is digitalization of life [1].

Both work on **discrete** information, represented by **sequences**. Required notion of alphabet A (examples: letters of natural language, chemical symbols-formulas, bits and bytes, numerical digits, DNA/RNA: U–T, keyboard).

[1] Chapter 1, Infobiotics

Information and Life

Which is the role of *information* in biological/living systems?

As a matter of fact, life emerged only when an efficient system of data processing was possible at a molecular level. Molecules are discrete structures built on atoms.

Life includes seven essential feature: birth, nutrition, growth, interaction, reproduction, death, evolution.
First 5 are individual instances of life, in time and space. Last 2 refer to forms of life of populations, a second level dynamics, providing the tree of life (i.e., genealogy of life species).

Sequences are linear (one dimensional) structures

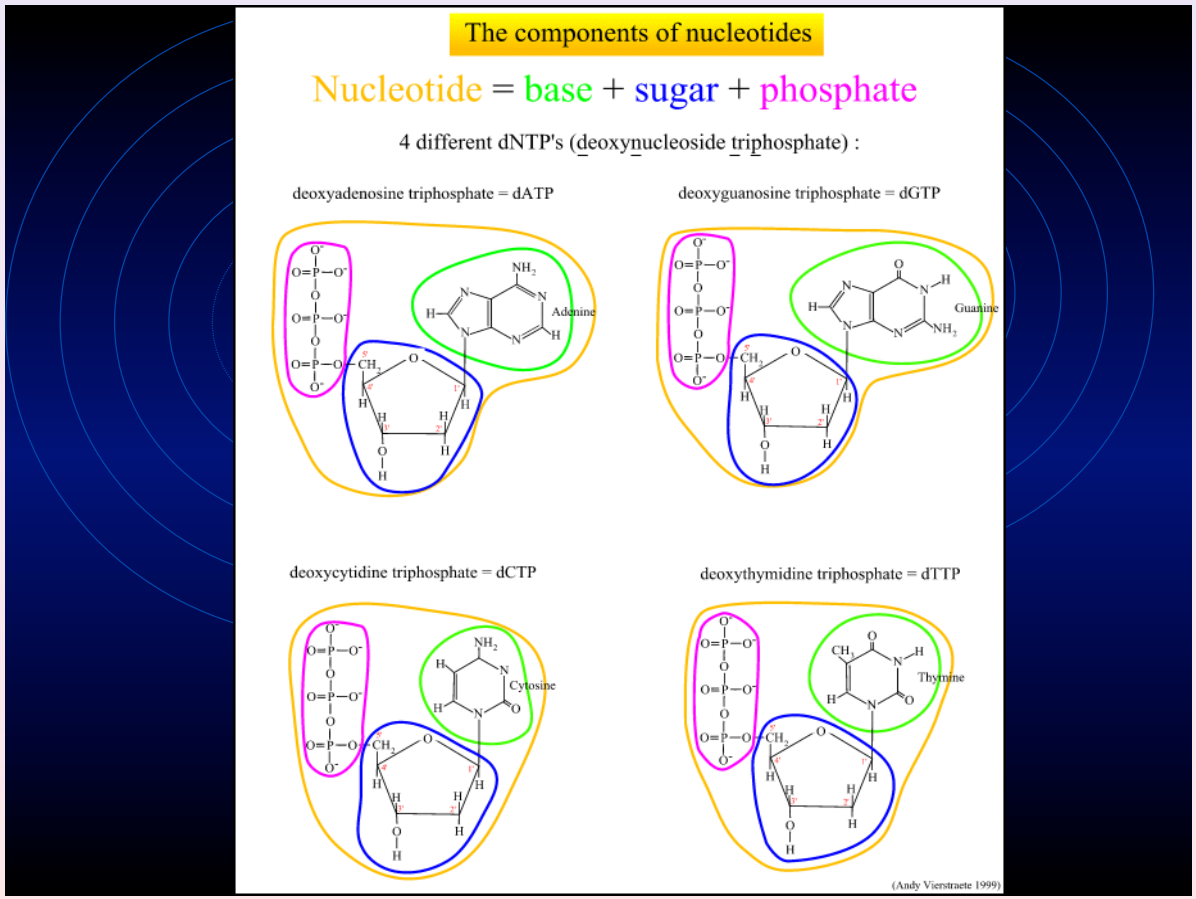
Information grows up exponentially with the sequence length. Lists and arrays. Set (no order, no repetition), multiset (repetition, no order), string (repetition, order).

Induction principle.
If $P(0)$ and $P(n-1) \Rightarrow P(n)$ for *any* n, then (P(n) for *all* $n \in \mathbb{N}$
Example: all cars have the same color.

Linear structures (biopolymers) are represented by sequences (of monomers) or strings. Subsequence more general than substring.

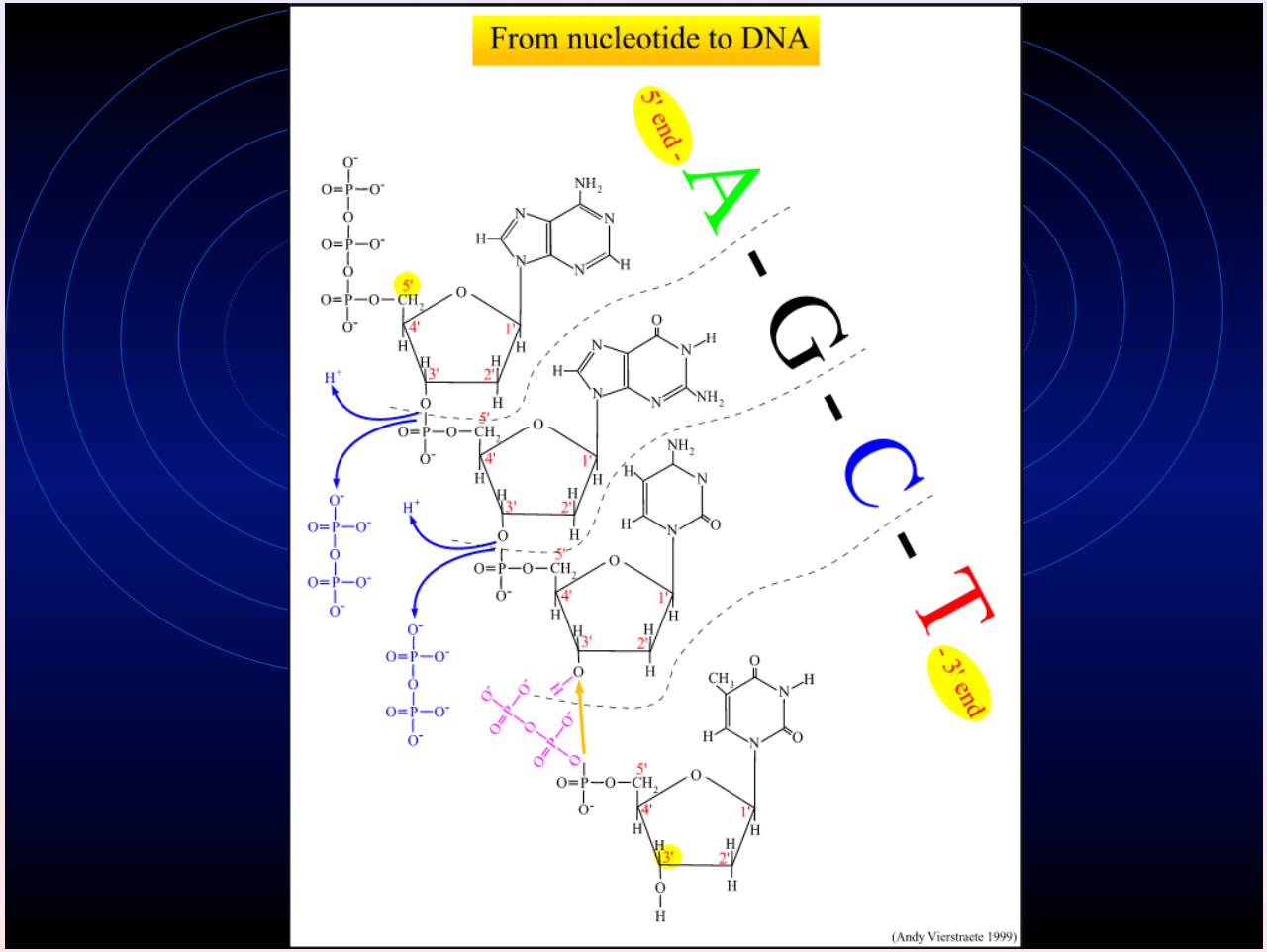
Nucleotide structure

Ribose = $C_5H_{10}O_5$, deossiribose = $C_5H_{10}O_4$ (no $2'$ oxigen).



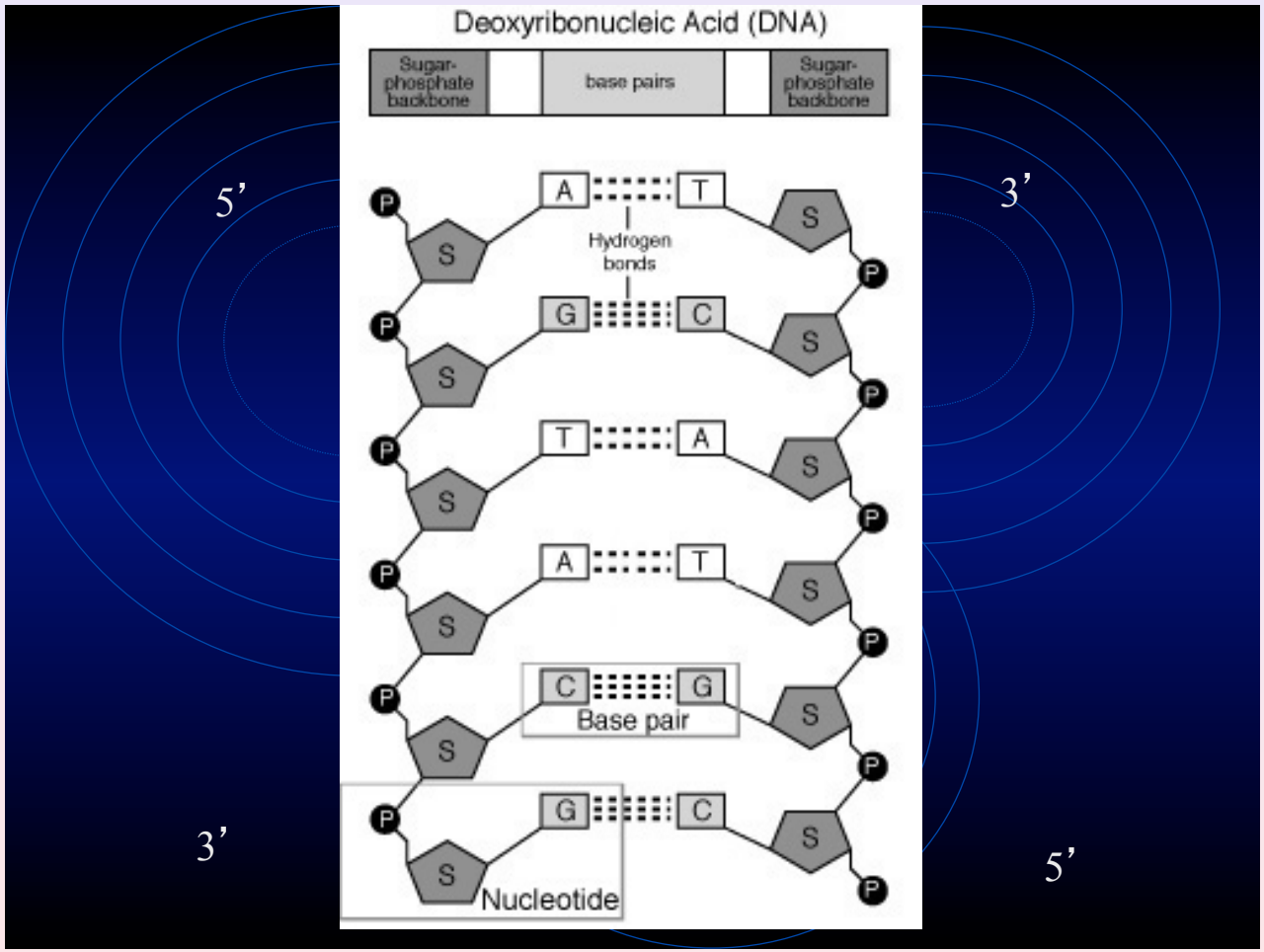
Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Sequence structure



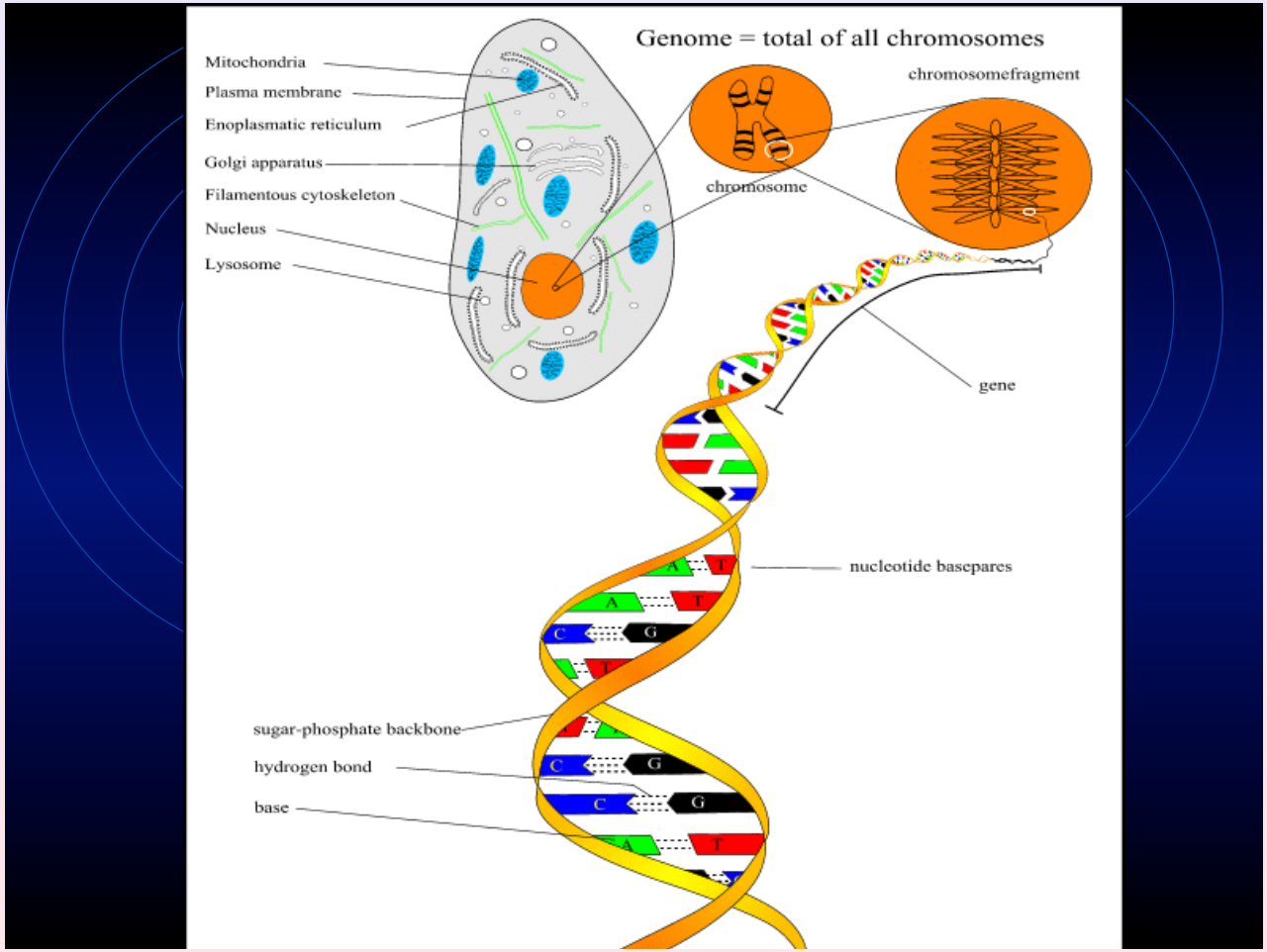
Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Sequence structure, by phosphodiester bonds



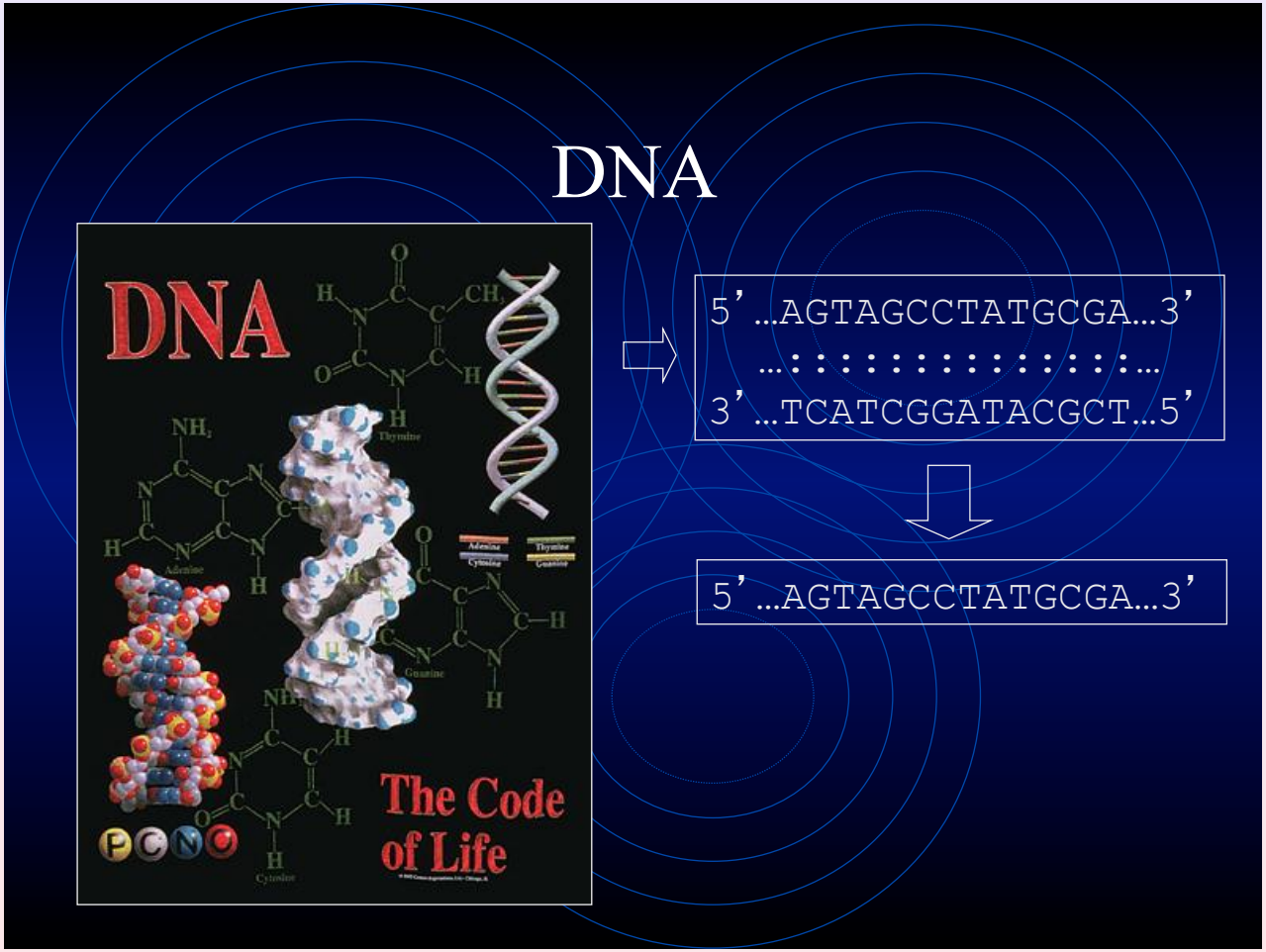
Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Double string structure, by hydrogen bonds



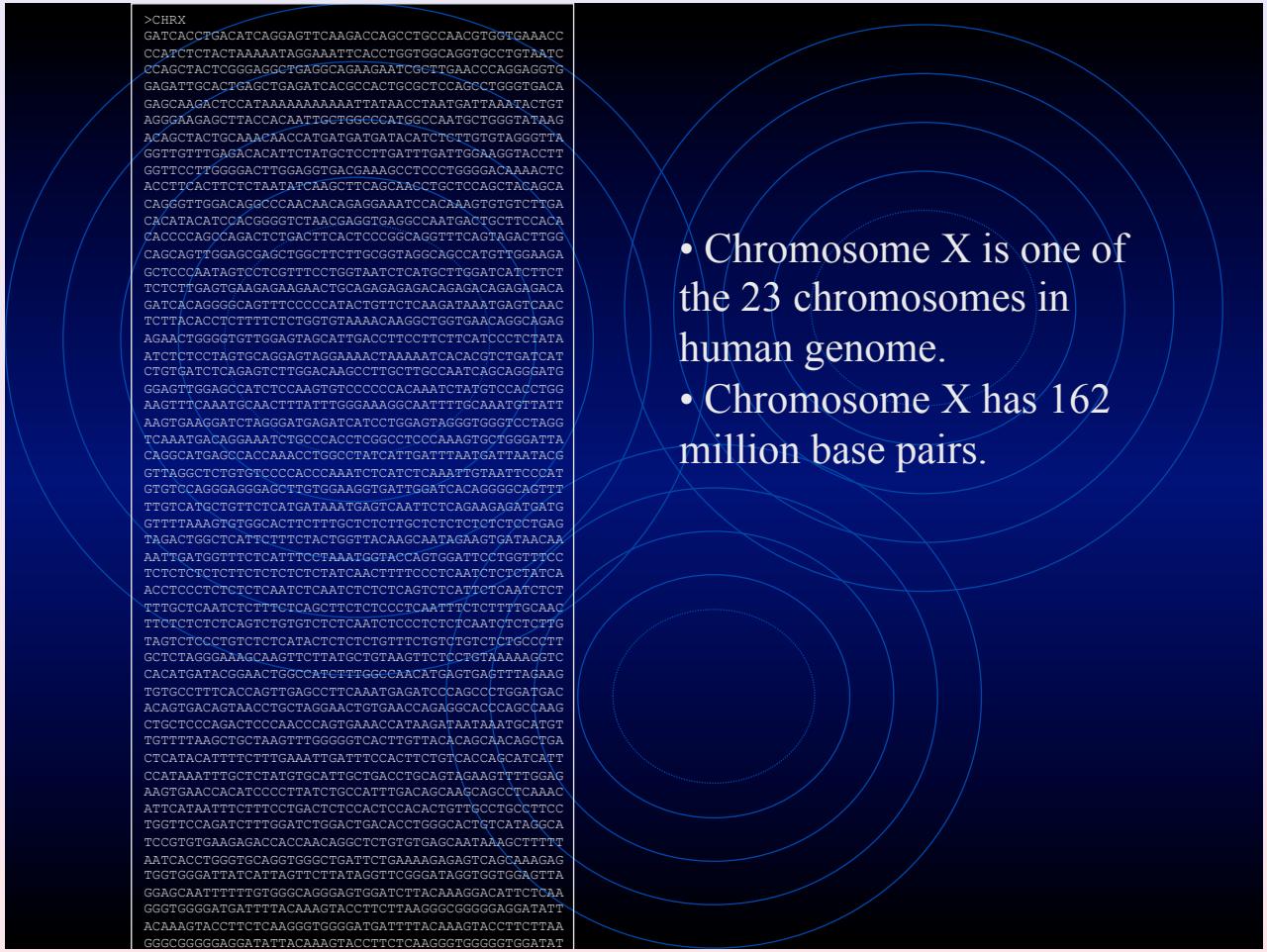
Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Secondary and Tertiary Structure



Slide courtesy of prof. Bin Ma, University of Waterloo, CA

DNA string



- Chromosome X is one of the 23 chromosomes in human genome.
- Chromosome X has 162 million base pairs.

Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Genome Sizes

Species	Size in bps
Amoeba dubia	670,000,000,000
Homo sapiens	3,400,000,000
Drosophila melanogaster	180,000,000
Mycoplasma genitalium	580,000
Human immunodeficiency virus type 1	9,750

Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Molecular and cellular systems

Life is born by complexification of molecular systems. Starting from RNA, then DNA, then (phospho)lipids, **polymers** (sequence of monomers) assemble to form a *protocell* (no growth, no reproduction). *Minimal cell* is an open system able to keep its own internal metabolism.

Formally, chiral and asymmetric molecules (such as DNA and phospholipids) may aggregate in either bilinear or spherical forms. Liposome float in water (inducing spherical forms).

Membranes enclose (inside-outside) and protect a reactor, group molecules close enough to react (by increasing their concentration), by catalysts or activators (special symbols).

Molecular and cellular systems

Starting from building blocks (polymers and phospholipids), they self assembly in (double) linear and spherical shapes (in water). Membranes induce reactions, then **reactions (and life) require water to start.**

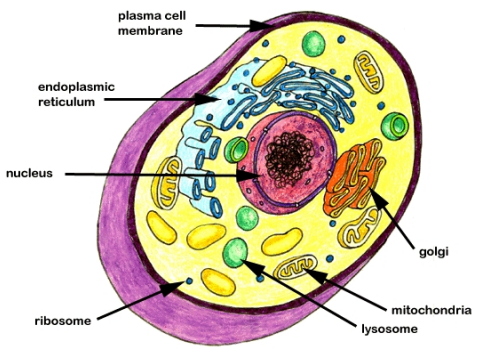
Bilinear molecules autoreplication follows Eigen paradox (error rate 1/100): longer replicator are both better replicator, with a greater error rate.

Replication requires reproduction (i.e., cell replication of both system and metabolic productions), so that best replicators are selected by evolution. **Efficient replication requires metabolism**.

Prokariotic versus eukariotic cells

LUCA (Least Universal Common Ancestor for modern cell, 3.8 billions years ago) is able to autoreproduce and to evolve. Prokariotic and eukariotic cells have differences: nucleus, aerobic function, obtained by fagocytosis of other organisms, presence of ATP.

Cells: complex systems of interactive components



- Two classifications of cell:
  - prokaryotic
  - eukaryotic
- Main actors:
  - membranes
  - proteins
  - DNA/RNA
  - ions, macromolecules,...
- Interaction networks:
  - metabolic pathways
  - signaling pathways
  - gene regulatory networks

Terminology

- The *genome* is an organism's complete set of DNA.
  - A bacterium contains about 600,000 DNA base pairs
  - Human and mouse genomes have some 3 billion
- Human genome has 24 distinct chromosomes (1-22, X and Y)
  - Each chromosome contains many *genes*
- **Gene**
  - basic physical and functional units of heredity.
  - Specific sequences of DNA bases that encode instructions on how to make *proteins*
- **Proteins**
  - Make up the cellular structure
  - Large, complex molecules made up of smaller subunits called *amino acids*

Slides courtesy of prof. Martin Amos, Manchester Metropolitan University, UK

Work of catalysis, transport, regulation, recognition, signaling..

Proteins

A gene is a substring of the DNA

- some genes are the "source code" of proteins

A protein is:

- a molecule
- structured as a string
- over an alphabet of twenty elements (amino acids)

Proteins have complex 3D structures related with their functions:

- Catalysis of chemical reactions (enzymes)
- Transport
- Structure
- .....

Basics of Life

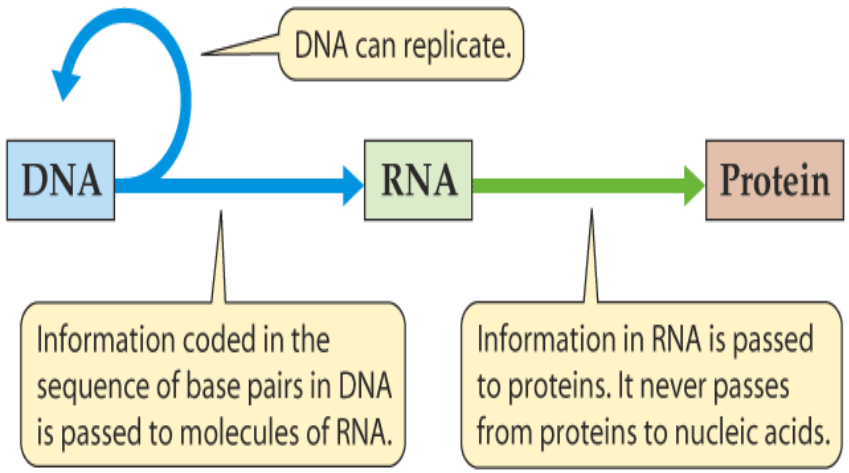
Cells store all information to replicate themselves (reproduction) in the genome. Almost every cell in human body contains same set of genes (and approximately the same genome). Genes are differently expressed in different tissues.

The genome is the operative system of the cell, which expresses one or the other type, just as computers are used for one or the other of their possible functions.

Totipotent (stem) cells are capable to differentiate into about 250 types of human cells.

From DNA to RNA

DNA, RNA and Information Flow



novel computation group

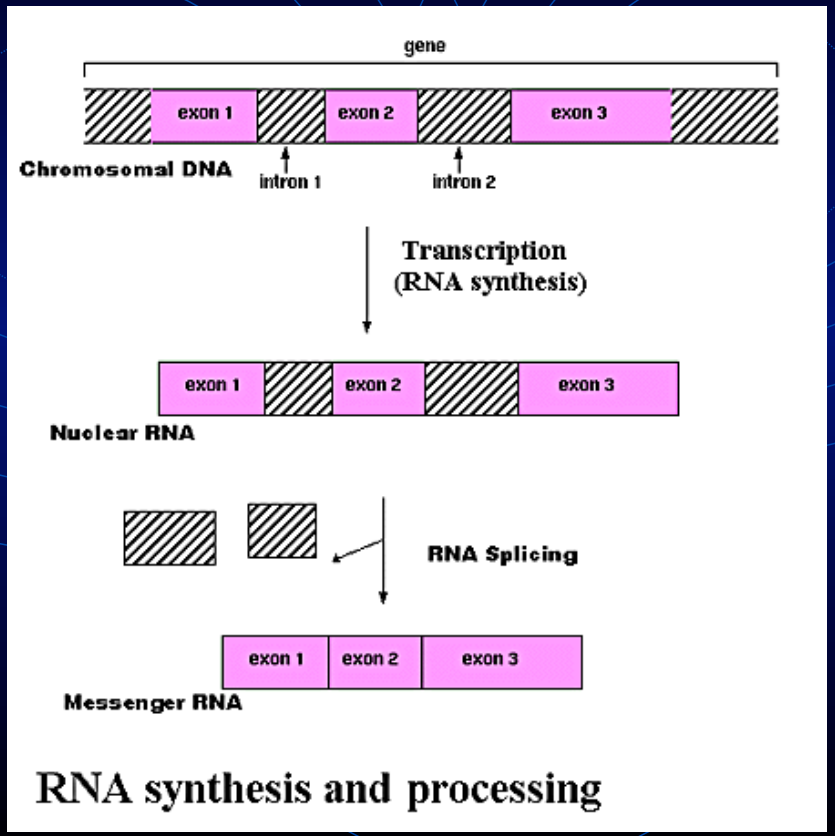
Slides courtesy of prof. Martin Amos, Manchester Metropolitan University, UK

Genes

- One gene encodes one* protein (or sometimes RNA).
- Like a program, it starts with start codon (e.g. ATG), then each three code one amino acid. Then a stop codon (e.g. TGA) signifies end of the gene.
- Genes are dense in prokaryotes and sparse in eukaryotes.
- In the middle of a eukaryotic gene, there are introns that are spliced out (as junk) after transcription. Good parts are called exons. This is the task of gene finding.

* Isoforms of one protein. [courtesy of prof. Bin Ma, Univ. of Waterloo, CA]

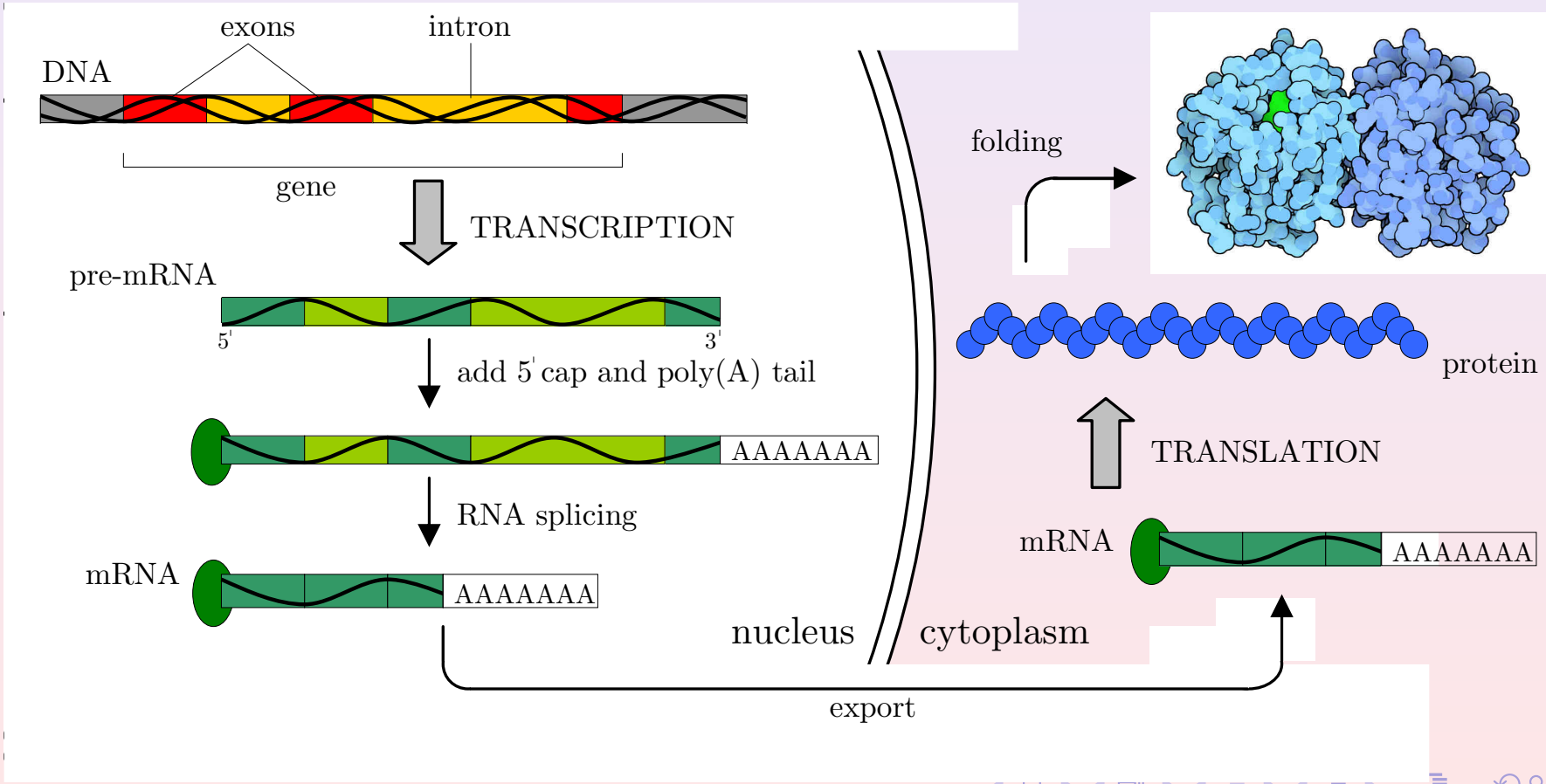
Introns and Exons



Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Central dogma (named by F. Crick)

Double helix wrapped around *histones*, assembled in 24 *chromosomes*: 3 billion bases in the cell nucleus (10^{-5}m).



Genetic Regulation

- **DNA contains the potential coding information for a vast range of possible proteins**
- **Gene expression is not a linear process**
- **Genes may require the product(s) of other gene(s) to in order to be expressed**
- **The product of one gene may turn off the expression of another gene**
- **The product of a gene can even effect its *own* expression (feedback)**

Slide courtesy of prof. Martin Amos, Manchester Metropolitan University, UK

RNAs

1640 genome-wide public datasets[2], for different types of cell, with a complete catalogue of:

- annotated human transcripts, identifying different types of RNAs: mRNA, (8800 short and 9600 at least 200b long) ncRNA, sRNA, rRNA, 32 (73-93b long) tRNAs, miRNA;
- functional elements, like promoters[3], millions of genetic switches, transcription factors, protein binding regions[4], transcription start sites (TSS), transcriptional repressors[5]

[2] regions of transcription and transcription factor association, chromatin accessibility and histone modification, by DNase

[3] `http://epd.vital-it.ch/`

[4] Es. CCCTC-binding factor

[5] Es, 15000-40000 binding sites of CTCF, 11-zinc finger protein

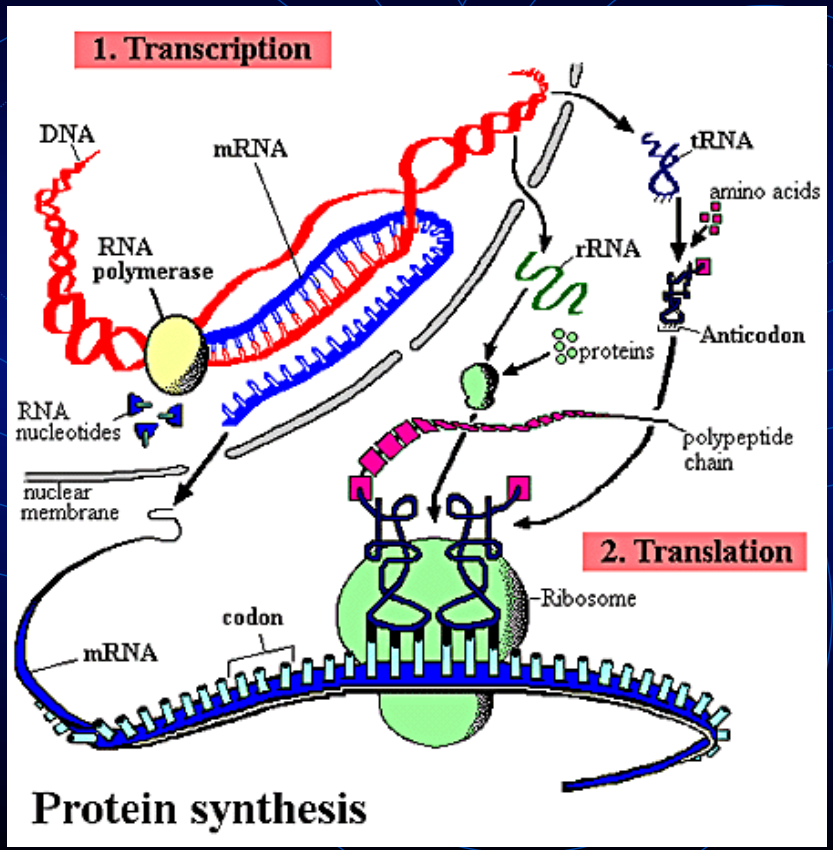
Combinatorial issues

Open problem (finding the *"genomic code"*): the relationship between regulatory elements and target genes.

Transcription factors bind at specific genomic locations in a combinatorial fashion to specify the on-and-off states of genes; the set of these binding events forms a cell regulatory network.

Many different genes may share the same transcription factors, but a *unique combination* of promoter sites (and transcription factors) works for a gene.

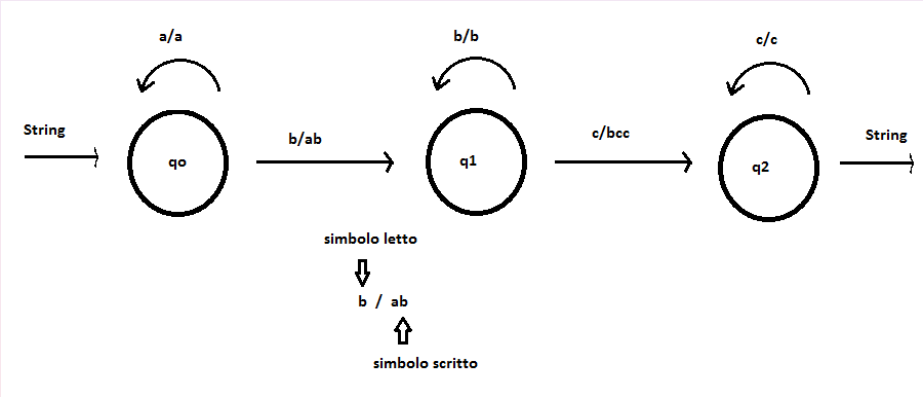
Protein synthesis



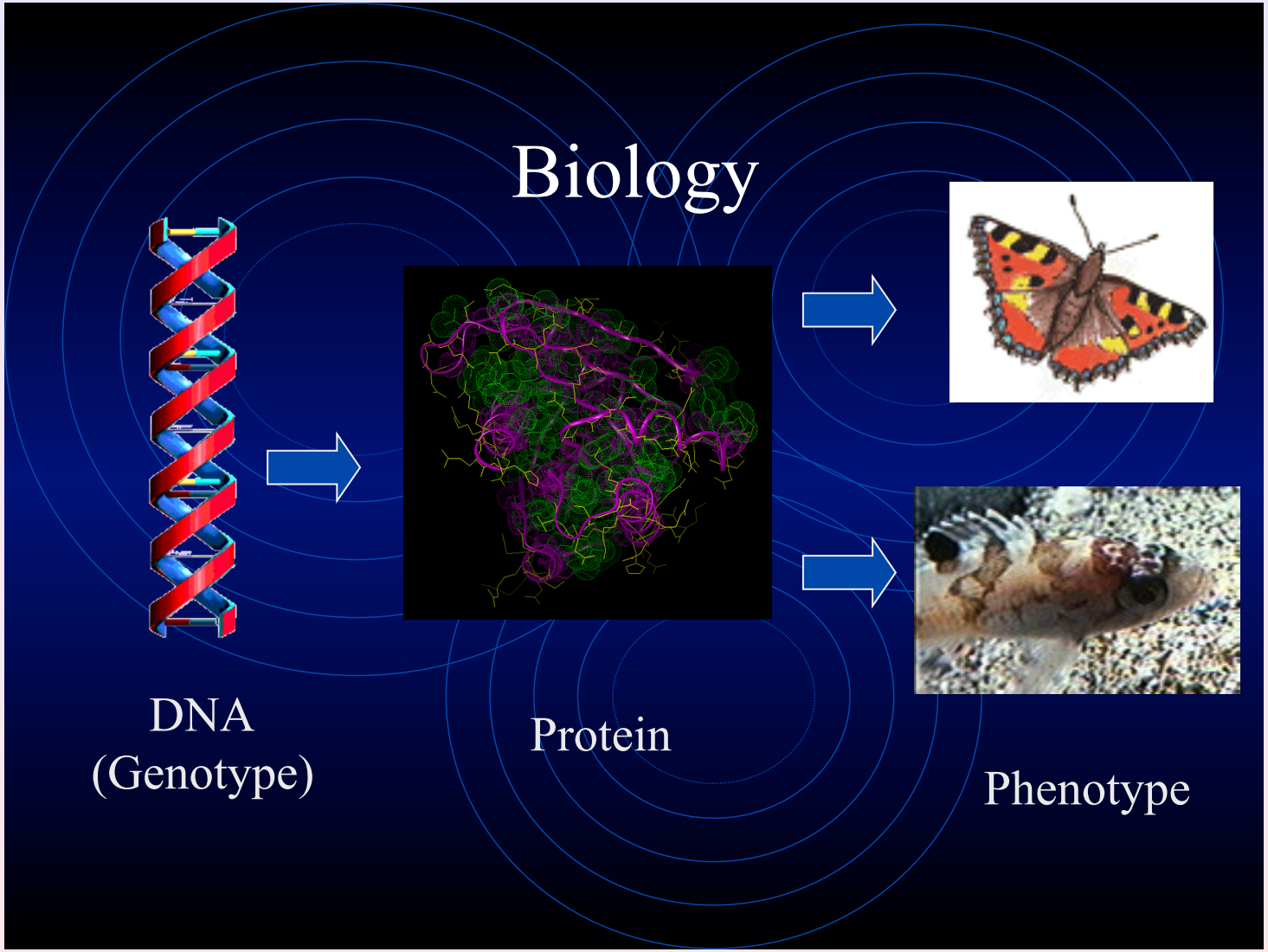
Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Rybosomial transducer

Ribosome is an iterated transducer (finite state automaton, able to replace strings – see Fig 4.2), with 25 transitions, and 16 final states. Simple example of a transducer generating the trisomatic language:



Since 25 transitions produce 21 outputs (aminoacids), then more than one transition gives a same output, *the genetic code (length 3, over 4 symbols) is not uniquely decodable*.



Slide courtesy of prof. Bin Ma, University of Waterloo, CA

Biological fundamentals

Fundamental dogma of biology (transcription and traslation): nucleic acid, gene, encoding sequence, aminoacid. Introns, exons, alternative splicing (by splicesomes) and protein isoforms.

Different types of RNA (e.g., pre-mRNA, mature-mRNA, RNA-tranfer). Concepts of automaton, tranducers, algorithm/program. Polymerases, rybosomes, and genetic code.

Two information levels between genes (about 20.000) and proteins (about 3.000): I-level (informational, genomic) and M-level (operational, metabolic).

Strings and Multisets

Polymers and membranes are molecular systems forming a cell-like environment. *Molecules and catalysts* react inside membranes, as multisets which are rewritten in parallel by rules (metabolic transformation), maybe applied with some strategy.

Chemical reactions are multiset trasformation by rewriting rules (according ti a *stoichiometric balance*).

Population is formalized by multiset (of molecules, membranes, cells, individuals, species).

Concepts of hyperset (power set), hypersequence (lexicographic order), hypermultiset (chemical formula).

Stoichiometric balance problem – (assignment I)

Given n multisets X_i and m multisets Y_j (over an alphabet), find minimum multiplicities (if they exist) $h_1, h_2, \ldots, h_n$, and $k_1, k_2, \ldots, k_m$ s.t.

$$h_1 X_1 + h_2 X_2 + \ldots + h_n X_n \rightarrow k_1 Y_1 + k_2 Y_2 + \ldots + k_m Y_m$$

Dalton (1803): matter of single objects is the same (both on left and right side).

Ex: in chlorophyll synthesis, light energy is converted in chemical energy of glucose: carbon dioxide and water produce glucose and oxigen (in mitocondria the opposite, aerobic reaction):

$$6CO_2 + 6H_2O \rightarrow C_6(H_2O)_6 + 6O_2$$

Findind the stoichiometric balance algorithm (ass I)

1. Start with numerical examples, for X_1 and X_2, Y_1 and Y_2;
2. Establish conditions for the multisets under which either stoichiometric values exist or they do not;
3. Find your algorithm;
4. Generalize it to $X_1, \ldots, X_n$, then to $Y_1, \ldots, Y_m$.

Tree and graph rewriting

(Membrane, phylogenetic) Trees for hierarchies. Rooted or not, parents, children, internal nodes, leaves, brothers, ancestors, descendant.

Trees represent a refinement (from strings) of order concept, and acyclic graphs. Defined by a *parent function*.

Graphs for (gene, protein) interactions. Nitrogen bases (see page 21) as graphs of atoms. Graphs over molecules (reactions), orcells (NNs). Cycle, (minimal, average, maximal) path, (in-/out) degree, diameter.

Formal representations

alphabet symbols $\rightsquigarrow$ molecules and monomers

sequences $\rightsquigarrow$ polymers

membrane $\rightsquigarrow$ spherical molecular aggregation

trees $\rightsquigarrow$ membrane structures

multisets $\rightsquigarrow$ molecule and membrane populations

multisets of rewriting rules $\rightsquigarrow$ chemical reactions (operation of sum and product by scalar on multisets)

Efficient realization of chemical reactions (graphs) requires compartimentalization (trees) and complex molecular systems (multisets). Those able to reproduce and evolve survive.

Formal and biological strings

Operations on formal strings (concatenation, iteration, λ, prefix, suffix, substring, length, reverse, reading, writing, permutation, editing: insertion, deletion, replace).

Operations on biological strings (complement, mir, length, writing, selective amplification, specific cuts, glue concatenation, pairing). Combinatorial pairing (Fig.2.14/15/16).

Differences: bilinear structure (no chirality), mobile/floating, deformable, "legible", chemical state.

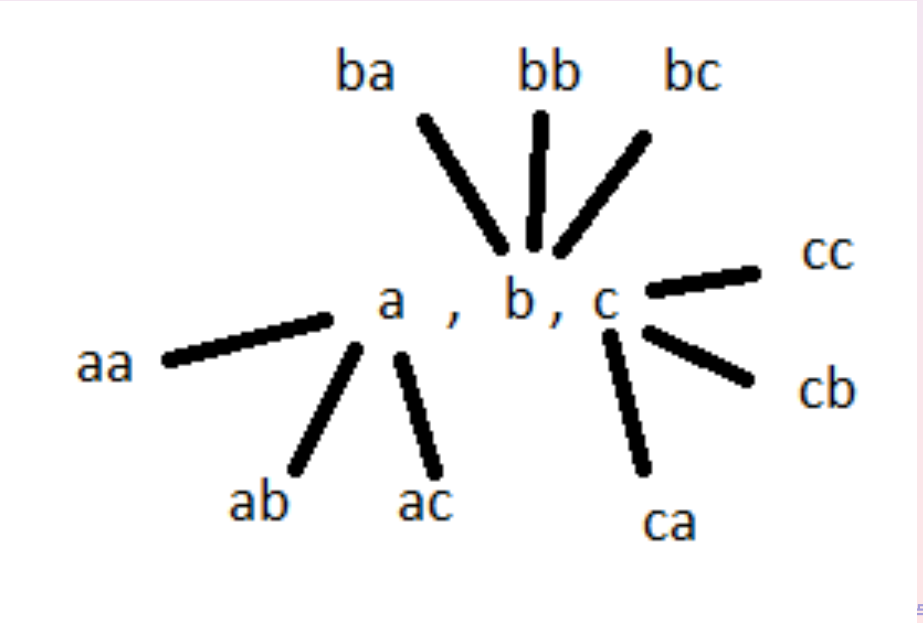
Why DNA? Stable, long-single-flexible /short-double-rigid, speedy and cheap synthesis, well known geometrical and thermodynamical properties, existence of enzymes and manipulation techniques.

Linguistic Universe

Over an alphabet A, we have a *free monoid* (algebraic strucuture with associative operation: concatenation)

$$A^\star = \cup_{n\in\mathbb{N}} A^n$$

with A^n defined by induction: : $A^0 = \{\lambda\}$, and $A^{n+1} = A \cdot A^n$
This is for example A^2:



Formal and biological languages

A language is an element of $\mathcal{P}(\mathcal{A}^\star)$, that is, a collection of strings.

Operations on languages (sum, intersection, difference, complement, concatenation, power, Kleene star).

$$L_1 \cdot L_2 = \{\alpha\beta \mid \alpha \in L_1, \beta \in L_2\}$$

Examples of languages (bipartite, bisomatic, trisomatic, duplicates, prefix, suffix, fix length codes).

Important investigations on formal languages as combinatorial discrete problems/functions.

SAT language/problem

Given a first order propositional formula ϕ, there are n Boolean variables connected by the three operators $\neg$, $\vee$, $\wedge$, which may be (efficiently) riorganized in a Boolean system of m *clauses*. Then, ϕ is equivalent to the conjunction of m clauses, each of

which is the disjunction of at most three *literals*, where each literal is a variable or its negation.

We say that $\phi \in$ 3-SAT(n, m) iff it is *SATisfiable*, meaning that there exists an assignment of truth values to the n variables making the formula true.

A 3-SAT(5,4) instance

Easy to solve formula:

$$\phi = (x_1 \vee \neg x_2 \vee x_4) \wedge (\neg x_2 \vee \neg x_3 \vee x_5) \wedge (x_1 \vee \neg x_4 \vee \neg x_5) \wedge (x_3 \vee \neg x_4 \vee \neg x_5)$$

Exact 3-SAT when clauses cannot be shorter than 3 literals.
3-SAT is NP complete problem, 2-SAT is a P problem.

Examer distributions – Assignment II

Over real genomic sequences of Pseudomonas or Mycoplasma (or over their genes):

(a) compute the number of segments containing *n* different examers (with n = 50, 100, 200, 500, 1000, 1500, 2000, 2500, 3000, 3200, 3500, 4096) and their length average. Diagram with 12 two-dimensional points for each sequence/organism.

(b) Check how minimal and maximal multiplicity of examers changes by elongating one segment over the genome.

Biological paradoxes

"Life is a paradoxal phenomenon, which evolves for solving paradoxes" [pag 171, Infobiotics].

Dualism gene-protein (double level of I-molecules and M-molecules) as a logical necessity of the autocatalytic nature of biochemical reactions in the cell.

Inventor's paradox [Pólya,2010]: The more ambitious plan may have more chances of success.

Life thinks at large – biological paradoxes

Sequence/Eigen paradox: shorter replicator is efficient for errors, longer replicator stably performs a more complex replication.
Efficient replication $\Rightarrow$ Metabolism, and reproduction.

Enzimatic/Rybosomic paradox: enzymes are special proteins, which catalize reactions, producing enzymes. Enzymes are hereditable from the I-level (of mather cell) and are then involved in the M-level. Necessity of an I-molecule/level, which means (auto)reproduction and heredity.

Metabolism $\Rightarrow$ replication $\Rightarrow$ metabolism/reproduction.

Life thinks at large – biological paradoxes

Evolutive paradox asks if life evolves to evolve (and fitness fuction evolves itself) or evolves according to a given Darwinian fitness – genetic drift helps to choose the first option.

Multiset paradox: multisets increase competition but make the system more stable (by redundance). It is solved by evolution, which select the most competitive multisets.

Morphogenetic paradox: where shapes are programmed in the cell?
Evo-devo (evolution, development) paradigm assumes the ovum existence, summa of totipotent cells (gametes have high recombinant power, due to the double chromosomal equipment) able to irreversibly differentiate, into several tissues.