MATHEMATICAL SLICES OF MOLECULAR BIOLOGY

A. CARBONE and M. GROMOV

Institut des Hautes Études Scientifiques
35, route de Chartres
91440 – Bures-sur-Yvette (France)

Janvier 2001

IHES/M/01/03

Mathematical slices of Molecular Biology

A. Carbone and M. Gromov

January 17, 2001

1 Introduction

Molecular biology and genomics promise to deliver a full set of structures needed to describe biological systems on the cellular level. Biomolecular techniques provide detailed (sometimes quantitative) information on the molecular functioning of the cell and allow systematic intervention into the cell functions. The scientific paradigm is shifting from "understanding the biological structure" to "tracking macromolecules", "manipulating and controlling the living cell", "creating artificial biomolecular structures", "modeling fragments of cell dynamics" and "identifying signatures of such fragments".

Mathematically speaking, the molecular biology encompasses the high dimensional set of parameters which projects to the lower dimensional space of the classical phenomenological biology. The behavior of the system within this low dimensional space rarely admits a self-contained quantitative description but on the molecular level such a description is feasible and entails great predictive power. The high dimensional space has a distinctive formal structure(s) which can be described, to some extent, in general terms without touching upon finer points. This makes the schema of molecular biology accessible to mathematicians.

In this article we shall try to transmit our enthusiasm for nano-level [1] cell biology and biomolecular techniques to the mathematical audience. Despite

[1] $10\text{Å} = 1nm = 10^{-3}\mu m = 10^{-9}m$ is a convenient scale in molecular biology, since most biomolecules are about 1-10 nanometers in diameter.

our limited (to say the least) knowledge of the cell we try to condense the picture into definite statements, not being afraid of saying something wrong or outrageous. We hope the reader will find this exposition sufficiently entertaining and provocative.

ACKNOWLEDGMENTS: Our fascination with DNA nanotechnology started with a lecture by Ned Seeman in 1996 in New York, and we had the pleasure of following his research along the years. Bud Mishra taught us the fundamentals of algorithmic and molecular aspects of genomics, and convinced us by his experience that a mathematician can productively work on biological problems. Sergei Mirkin projected to us the conceptual beauty of designing subtle experiments and encouraged us to think about molecular biology from a mathematical angle. Alexander Gorban demonstrated to us possibilities and limitations of mathematical and computational modeling of biophysical systems. We want to thank the speakers of our seminar for the inspiring expositions of their experimental and theoretical work. Our special thanks go to François Képès who spent hours generously sharing with us his vaste knowledge of molecular biology and his ideas on structural organisation of biological systems. He and Ned Seeman suggested many improvements to the manuscript, and Christophe Soulé helped us to clarify several points.

2 Crick's dogma

The ancestral memory and the program running the cell are encoded in *DNA*, a long chain of 4 small molecules. In a living cell, some segments of DNA are copied or, as molecular biologists say, *transcribed*, by means of *proteins* into shorter chains of similar molecules. These chains, called *mRNA*'s, are further *translated* to *polypeptide chains* (proteins) made of 20 amino-acids (another class of basic small molecules). In the course of translation (or soon thereafter) proteins fold into compact three dimensional conformations, and mutually interacting, make up the architecture (mainly by self-assembly) and the dynamics (e.g. catalytic activity) of the cell. The production of mRNA and proteins is accompanied by a continuous process of degradation, which is less understood (especially for proteins) than the synthesis. This schematic picture, properly annotated to be truly correct, will be further refered to as *Crick's dogma.*

3 Static and dynamic structures

The dynamics of the cell is a continuous flow of small molecules channeled by the interaction with macromolecules: DNA, RNA and proteins. The behavior of small molecules obeys the statistical rules of chemical kinetics, where, in particular, the rate of reaction is proportional to some powers (usually small integers) of concentrations [2] of the reactants. Sizeable amounts of macromolecules might be described in a similar way. At the *mesoscale* however, macromolecules and macromolecular complexes appear as individuals [3], tiny mechanical contraptions, handling and sheparding small molecules: enzymes controlling metabolic pathways and being themselves switched on and off by small molecules, RNA polymerase synthesizing RNA out of nucleotides using DNA as a template, ribosomes synthesizing proteins from amino-acids with the help of tRNA, proteasomes selectively degrading proteins, etc. Some components of these machines can be used outside the cell for directing specific mesochemical processes (i.e. chemistry on the mesoscale such as PCR and protein engineering) and the main challenge is to create new mesochemical devices, comparable in structural complexity and specificity to those used by the cell. The design and function of some mesochemical machines are presented later in the paper.

DNA. This is a long (sometimes very long!) *word* in the four letters A, T, C and G. These letters name the four nucleotides constituting DNA. The molecule of each nucleotide is built out of the *sugar-phosphate group* and the *base* attached to it. The sugar-phosphate group is the same for all four nucleotides, while the specificity is due to differences between the bases: A (Adenine), T (Thymine), C (Cytosine) and G (Guanine). (See Figs. 1 and 4.)

The sugar-phosphate group is naturally *polarized.* Each nucleotide can be *covalently* bound to another one with a phosphate group bridging between

[2]If the channeling and the compartmentalization effects induced by macromolecules reach the nanoscale, then the averaging implicit in the notion of "concentration" and "ideal kinetics" becomes questionable. Also, some small molecules, important for cell regulation, appear in low numbers (e.g. 10) in small prokaryotic cells and statistics should be applied more carefully. In the presence of enzymes, the polylinearity of kinetics may break down, as it happens in metabolic pathways.

[3]The individuality of a large molecule emerges in the statistical ensemble of atoms constituting this molecule.

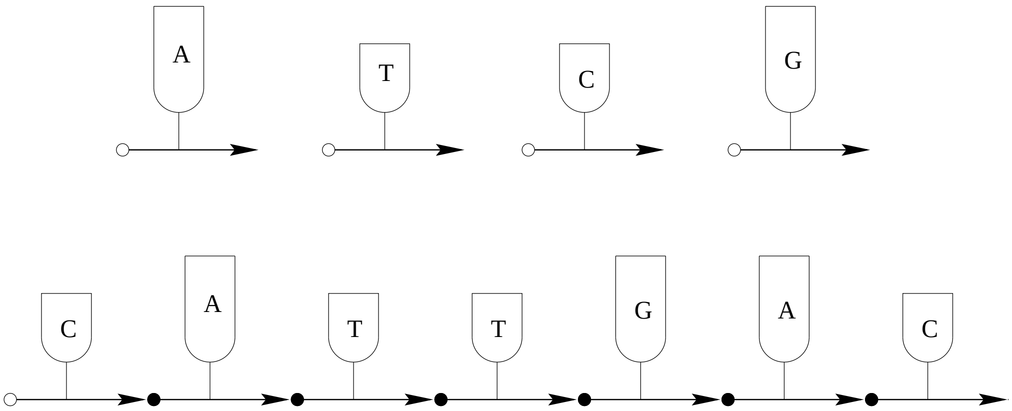

Figure 1: The four nucleotides (on the top) and a polynucleotide chain (on the bottom). The (sequence of) arrows and circles represent the *sugar-phosphate backbone*, while the palettes correspond to *bases*. The small *white* circle represents the phosphate group which disactivates (by loosing part of it) in the course of polymerization, and then it is depicted with a small *black* circle. If we break a DNA sequence into individual nucleotides, these will be disactivated and will be unable to make a chain again without an import of energy.

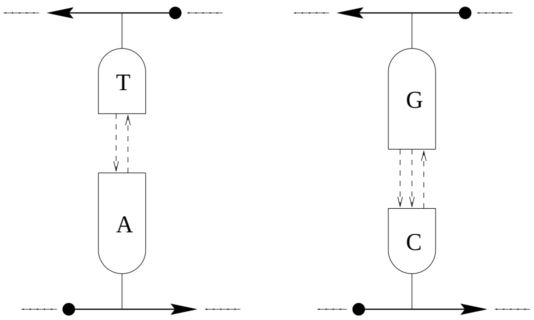

Figure 2: Hydrogen bonds between nucleotides (dotted arrows). Observe that nucleotides are antiparallel under the hydrogen bonding and that the AT interaction is weaker than the CG interaction as produced by two hydrogen bonds in AT and three bonds in CG. Each dotted arrow indicates a bond issuing from an hydrogen atom of the base, as seen in Fig. 4.

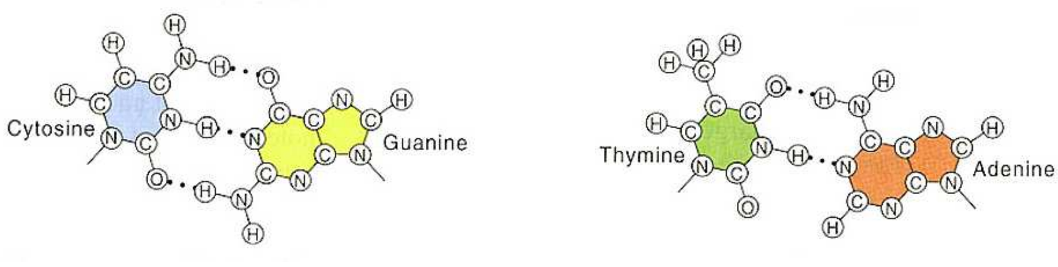

Figure 3: A chemical representation of the Watson-Crick complementary bases.

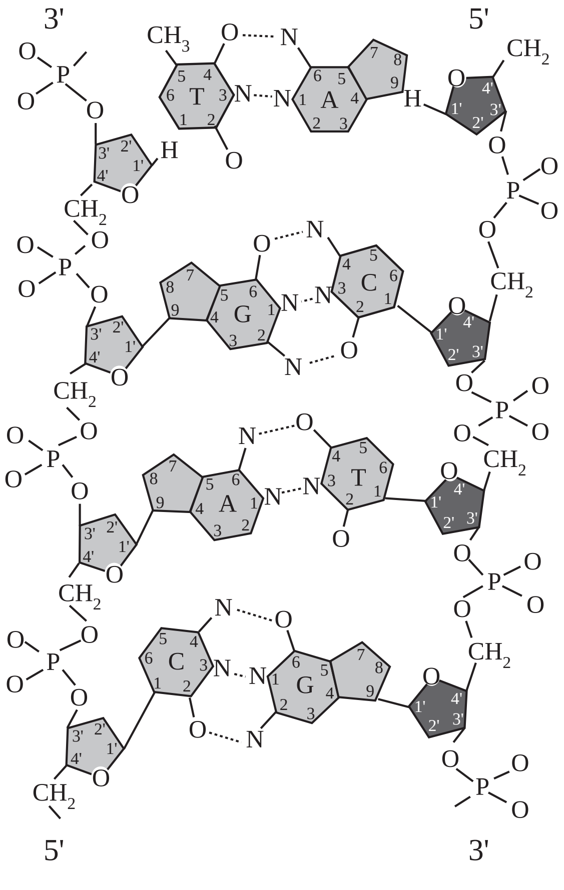

Figure 4: A chemical representation of a double stranded DNA. The two sugar-phosphate backbones are bridged by coupled base pairs.

two consecutive sugars, and this bond agrees with the polarization, as it is schematically shown in Fig. 1, with a more realistic picture displayed in Fig. 4. This process can be repeated almost indefinitely and produces long chains of nucleotides called *polynucleotides* or *single stranded DNA*. Short polynucleotides composed of $\lesssim 100$ nucleotides are called *oligonucleotides*. They can be nowadays synthesized chemically according to a given specification of letters.

Crick-Watson complementarity. The functioning of DNA, and thus all life on earth, depends on the particular affinity between the bases A to T, and C to G, which comes from another kind of chemical bonding between nucleotides called *hydrogen bonds*, which are about 10 times weaker than the covalent bonding described above. See table in Fig. 6.

If we throw into a solution many copies of the nucleotides A, T, C and G, then soon they will predominantly appear in complementary pairs AT and CG. [4] See Fig. 2. (The covalent sugar-phosphate bonds between nucleotides do *not* form spontaneously.) Furthermore, suppose that we have a solution of various polynucleotides in a tube. Whenever two inverse complementary subwords appear in two (possibly the same, but not adjacent in the chain) *poly*nucleotides, these eventually come close together and attach to each other along these subwords [5]. (See Fig. 7.) The longer these words, the stronger this attachment will be.

As molecules are in thermal motion, the binding along complementary words, called *hybridization*, appears and disappears. There is a continuous competition between hybridization of different complementary subwords in the polynucleotides and there is no satisfactory quantitative theory predicting the statistical distribution of hybridization. However, in accordance with one's intuition, the presence of pairs of long complementary subwords makes

[4] In a pot, every base can pair with every other base, including itself; G and C will preferentially pair together, and there are suggestions that A and T also will. The A-T pairing however may not be Watson-Crick, as there are three other types of A-T pairs, reverse Watson-Crick, and two other (Hoogsteen) mutually reversed pairings, which in the absence of a backbone seem to be equally likely. There are more than 30 possible hydrogen bound pairs between bases and also several base triplets. The latter occur in nature, but in RNA (some tRNA) rather than in DNA.

[5] The specificity of AT and GC binding is by far more pronounced in polynucleotides than for monomers. This is due, to a large extent, to the size matching: A and G are roughly twice as large than T and C, so that the pairs AT and CG fit nicely into the DNA double helix.

hydrogen atom		1.008Da
carbon C^{12}		$=_{def}$12Da
nitrogen, oxygen atoms		14Da,16Da
sulfur, phosphorus atoms		32.06Da,30.97Da
covalent bonds (distances between nuclei)	0.75-2.3Å	
hydrogen bonds	1-2Å	
van der Waals radius	2-3Å	
water molecule	3-4Å	18Da
sugars, amino-acids, nucleotides	0.5-1nm	150-500Da
globular proteins	2-10nm	$5 \cdot 10^3$-$5 \cdot 10^5$Da
ribosomes	25-30nm	2.5-4.5MDa
viruses	26-60nm	3-50MDa
bacteria	0.5-5μm	$5 \cdot 10^3$-$5 \cdot 10^6$MDa
mitochondria	2μm	10^5-10^6MDa
cell nucleus	3-10μm	10^6-$5 \cdot 10^7$MDa
animal cells	10-30μm	$5 \cdot 10^7$-$5 \cdot 10^9$MDa
plant cells	10-100μm	$5 \cdot 10^7$-10^{11}MDa
DNA in a human cell	2m	$5 \cdot 10^6$MDa
DNA in a human body	$10^{14}m$	200-400g

Figure 5: Table of linear scales and masses. All the numbers in the table are *approximate* values. Da stands for 1 dalton $\approx 1.66054 \times 10^{-24}$g $= 1/N_A$, where N_A is the Avogadro number 6.0221367×10^{23}. The weights of atoms are averaged over the distribution of the isotopes in the biosphere. The size of an atom or of a molecule refers to the radius of the chemical force it exerts in a particular class of chemical interactions (covalent, ionic, van der Waals, etc.). Thus, the sizes of atoms can be regarded as distances between nuclei of covalently bound atoms; these range between 0.75-2.3Å for atoms present in the cell as ions or in molecules. Masses of macromolecules, ribosomes and viruses refer to the "dry weight", i.e. without water, while larger unities are weighted with water, which makes $\approx 70\%$ of the whole weight. DNA make about 1% of the total weight in bacteria and about 0.25% in mammalian cells.

average kinetic energy of thermal motion per molecule at room temperature (25°C $\approx 293°K$)	0.9
energy of a hydrogen bond	1-5
approximate energy of covalent phosphate bond between nucleotides	50
approximate energy of $C = C$ bond	200

Figure 6: Energy table. The common chemical unit of energy is $kcal/mol$, where $1mol$ is the amount of substance containing $N_A = 6.0221367 \times 10^{23}$ molecules. For example, one mole of water makes about $18g$, since the molecule H_2O weights $\approx 18Da$. Remind that $1kcal = 4,184J$, where 1 joule (J) equals the kinetic energy of linear motion of $1kg$ of matter with the speed of $1m/sec$. The average thermal energy of linear motion of molecules at the temperature T measured in Kelvins (K) ($T_{Kelvin} \approx T_{Celsius} + 273$), equals $\frac{3}{2}kT$ joules, where $k = 1.380658 \times 10^{-23} JK^{-1}$ is the Boltzmann constant. At the room temperature $T = 298K$, the thermal energy is $\frac{3}{2}kTN_A/4,184$ makes $0.8882855 \approx 0.9kcal/mol$ as indicated in the table. The other energies in the table are also expressed in $kcal/mol$. The velocity V of a particle of mass m kilograms with the kinetic energy T kelvins, is $V = \sqrt{\frac{3kT}{m}}$. For example, the water molecule $\approx 18Da \approx 3 \cdot 10^{-26}kg$, has (quadratic) average speed $V = 640m/sec$ at room temperature.

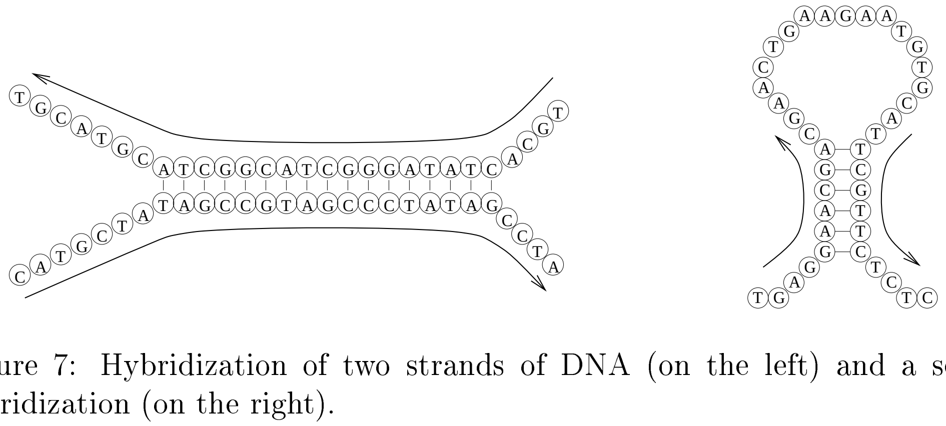

Figure 7: Hybridization of two strands of DNA (on the left) and a self-hybridization (on the right).

Species	Haploid genome size	Haploid chromosome number
E. coli (bacteria)	4,640kb	1 (circular)
S. cerevisae (yeast)	12,050kb	16
D. discoideum (slime mold)	70Mb	7
C. elegans (worm)	100Mb	11/12
A. thaliana (weed)	130Mb	5
D. melanogaster (fruit fly)	170Mb	4
Gallus domesticus (chicken)	1200Mb	39
M. musculus (mouse)	3Gb	20
X. laevis (toad)	3Gb	18
H. sapiens (human)	3Gb	23
Zea mays (maize)	5Gb	10
Allium cepa (onion)	15Gb	8

Figure 8: The length of the genome in different species. The symbol kb denotes one thousand base pairs, Mb stands for 1000kb and Gb denotes 1000Mb. The length of the genome only roughly corresponds to the "complexity of the organism". The organization of the genome into chromosomes does *not* reflect the length of the genome.

the corresponding nucleotides stick along these subwords and stay together for a long time, at least at the room temperature. As temperature reaches roughly 100°C, the denaturation (i.e. the breaking of the hydrogen bonds between the complementary nucleotides) becomes prevalent.

Association as well as dissociation of neighboring hydrogen bonds between complementary words are a positively correlated process leading to a "zipping" effect: hybridization starts at the ends of two complementary words and then persists at a high speed, or conversely, disengagement of two words unzips from an end on. The (high dimensional) landscape of the energy describing this interaction appears as a multiscale network of rivulets merging into the river corresponding to the zipping/unzipping mechanism.

The genetic information in a cell is encoded in a *double stranded DNA*, consisting of two complementary strands of equal length held together (quite strongly) by the hydrogen bonds. Thus, the unit of information is a pair of complementary nucleotides, commonly abbreviated to *bp* for *base pairs*. Fig. 8 gives approximate numbers of bps for genomes of various species.

In *eukaryotic* (i.e. "higher", from yeast on) organisms [6] the genome is organized into several units, called *chromosomes*, which are connected components or separate words of DNA, ranging in number from a few up to several dozens. (See Fig. 8.) In bacteria there is typically a unique circular DNA. Besides, there is extra genomic information contained in rather short circular DNA present in both eukaryotic cells (in organelles) and in bacteria (as plasmids).

A DNA word within a chromosome is intricately folded and spatially (biologists say "*topologically*") organized in a highly structured way, so that a 4cm DNA for example fits into a chromosome body $\approx 3 - 10\mu m$ long and $\approx 1\mu m$ thick. In bacteria, the circular DNA is also compactified (by *supercoiling*) but in a much simpler way than for eukaryotes.

The redundancy of having two strands instead of one, pays off in a variety of ways:

> **replication**: the replication mechanism using double stranded DNA invented by nature seems to represent the simplest possible logical solution to the von Neumann problem of self-replicating automata, beating von Neumann implementations by several orders of magnitude in simplicity and mathematical elegance. (Expressing the idea of self-replication in *general* mathematical terms remains an open problem. The existing models are essentially limited to the Ulam-von Neumann framework of cellular automata.)
>
> **repair**: doubling of information allows the cell to employ various error correction mechanisms to ensure the fidelity of replication.
>
> **stability**: double stranded DNA is by far mechanically stronger and chemically more resistant than the single stranded DNA.

Why four letters $(A, T, C$ *and* $G)$ *rather than two?* One can think of a single stranded DNA as a word in the free group F_2 generated by A, G with

[6] Eukaryotic cells are about 10 times larger than *prokaryotic* (i.e. bacterial) ones, and they have a fine internal structure: a nucleus which contains DNA, a skeleton, semi-independent membrane-bounded entities called organelles, etc.

$T = A^{-1}$ and $C = G^{-1}$, where hybridization corresponds to cancellation between reciprocal words. Four letters help to avoid excessive cancellation (manifested by undesirable hybridization): two letters would make an abelian (infinite cyclic) group.

Partial hybridization has more bio-chemical significance for RNA than for DNA, as the latter appears in two strands "cancelling" each other and not leading directly to sophisticated geometric patterns. The arrangement of the ribosomal RNA (and, to a lesser extent, of transfer RNA) is reminiscent in shape of the van Kampen diagrams familiar to combinatorial group theorists. See Figs. 11 and 10. (One wonders whether the distribution of RNA/DNA words in the abelianized group $F_2/[F_2, F_2] = \mathbb{Z}^2$ or in the higher nilpotent quotients of the free group bear any biological significance.)

RNA. Chemically, RNA is a twin sister of single stranded DNA: it has a slightly different sugar-phosphate backbone and a small modification of the four bases. It keeps the same A, C, G bases while T (Thymine) is replaced by its chemical relative U (Uracil).

Unlike DNA, RNA appears in cells in millions disconnected segments 100-1,000,000 nucleotides long. RNA's are classified according to the function that they perform in the cell. *Messenger* RNA's, in short mRNA, serve as go-betweens DNA and proteins. *Transfer* RNA, or tRNA, and *ribosomal* RNA, or rRNA, are incorporated into the machinery implementing the information carried by mRNA into synthesis of proteins. mRNA's are 100-10,000 nucleotides long, tRNA's are about 100 and rRNA are 2,000-5,000. Some of these RNA's are produced from short lived intermediates, called *pre-RNA*'s, which can be up to 10^6 nucleotides long. In a rapidly growing mammalian cell, the 80% of the total RNA is rRNA, the 15% is tRNA, while mRNA make a small portion of the total RNA.

Why RNA at all? Could one design a cell where all functions of RNA are performed by segments of single stranded DNA? The prevalent point of view suggests that RNA appeared at the early stages of life with DNA and proteins coming much later. This is witnessed by the presence of RNA in several ancient machineries, including ribosomes and spliceosomes [7]. It was

[7]See Section 3 for the definition of ribosome. Spliceosomes are ribonucleoprotein complexes roughly the size of ribosomes. They assist transformation of pre-mRNA to mRNA in eukariotes.

discovered relatively recently that certain conformations of RNA (i.e. specific three dimensional shapes it takes under the influence of weak interactions between its subsegments) display (auto-)catalytic properties similar to those of certain enzymes (proteins), which can conceivably assist the replication process. Thus RNA may appear in two distinct roles: a keeper of information and a worker executing this information for the purpose of replication. Structurally, information is organized in a straightforward way as a string of letters coding nucleotides, while the catalytic activity depends on a three dimensional arrangement of the polynucleotide chain.

How does the linear chain of nucleotides acquire a specific three dimensional shape? Its behavior in solution follows paths that can be rather faithfully described by a system of stochastic differential equations where the underlying potential U incorporates the interaction between various molecules constituting the chain and those in the solution, and where the major contribution to U comes from the hydrogen bonds responsible for the complementarity. A computationally feasible solution to the system runs into a multitude of serious problems (similar, but possibly less severe, than those for protein folding) where the major difficulty is the exponential multitude of the local minima of U. Some of these can be accounted for by the combinatorial patterns of matching between complementary words. This matching determines what is called, the *secondary structure* of RNA, and there are heuristic algorithms for its determination. Deriving the final three dimensional shape from the secondary structure remains an unsolved (to some extent mathematical) problem.

Transcription is a transformation of specific segments of DNA into single strands of RNA. There are certain proteins, called *transcription factors*, that choose particular segments of DNA, which then serve as templates for the production of the complementary segments of RNA. The synthesis is performed with the help of another protein, the *RNA polymerase* (which binds to one of the strands of the double stranded DNA), with an average rate of 60 nucleotides per second in E. coli. (This number may be different for other organisms.)

The resulting RNAs are further modified (*cleaved* and /or *spliced*) with the help of various enzymes. For example, in eukaryotic cells, pre-mRNA is spliced by cutting away several (possibly large) subsegments (*introns*) and by sewing (*ligating*) together the remaining segments (*exons*).

The splicing process is not unique (even the division into introns and exons is not canonical) but depends on a particular biochemical (physiological) state of the cell. Also, in viruses and prokaryotes, one segment of DNA may be transcribed into to several disjoint segments of RNA. The "edited" RNA is called messenger RNA, or mRNA. The length of mRNA produced in a single cell may vary from few hundreds to 10,000 (and sometimes more) nucleotides.

Pre-mRNA folds in the course of transcription and this folding plays a role in the editing of pre-mRNA to mRNA. The co-transcriptional folding follows a different path from that of free molecules in solution, due to the constraints imposed by the transcription process. Roughly speaking, instead of a single system we have a sequence of "correlated" systems of differential equations time-indexed by the number of nucleotides transcribed at a given moment, and depending on the composition of the synthesized segment. The solution of these equations must be time-correlated with the time-indexing of these equations.

Definition of a gene. Originally a gene was understood as an abstract unit of heredity but there is no consensus nowadays on which structure or biological function corresponds to such a unit on a molecular level.

As mathematicians we are less sensitive to the distinction between (spatial) "structures" and "functions", as the latter appears as structures seen in the space-time. With this in mind, we define[8] a gene as a segment of DNA together with a transformation to a segment(s) of mRNA (or other functional RNAs such as tRNA, or rRNA). This definition is supposed to capture two phenomena: *alternative splicing* in eukaryotic cells, where the same segment of DNA may lead to the production of different mRNA depending on a particular global state of the cell or on what happening in the vicinity of the transcription site; *overlapping genes*, where different segments of RNA are produced from overlapping segments of DNA, as it happens in viruses and prokaryotes.

The first level of the structure of the genome consists of the sequences

[8]In biology, a *definition* is supposed to isolate a pronounced phenomenon but not necessarily capture it completely. For instance, one cannot give a definition of a gene being both concise and exhaustive. Besides, the mathematical standpoint suggests a definition of an object along with the full category of related objects (and morphisms). Here the relevant category can be made of genetic networks, so that a definition of gene would include the regulation of the expression of the gene.

Species	Genome size	% coding DNA
E. coli (bacterium)	4,640kb	100
S. cerevisae (yeast)	12,050kb	70
C. elegans (worm)	100Mb	25
D. melanogaster (fruit fly)	170Mb	33
H. sapiens (human)	3Gb	15
Protopterus (lungfish)	140Gb	0.7
A. thaliana (plant)	130Mb	31
Fritillaria (plant)	130Gb	0.02

Figure 9: The length of the genome in different species compared with its coding part, where the latter is a very rough and disputed estimate for large genomes.

of letters A, T, C, G with distinguished segments supporting genes. Parts of these are transcribed to mRNA (with the introns spliced away) and other parts, called *regulatory region(s)*, are responsible on when and how the transcription takes place. (A regulatory region may be separated from the transcribed part by a long stretch of DNA possibly containing other regulatory and transcribed segments. Sometimes such a region is respectfully called *regulatory gene.*)

The intergenic space constitutes the bulk of DNA in most eukaryotes and it is unrespectfully referred to as *junk DNA*; a similar disrespect is extended to introns. See Fig. 9.

The sequences of letters constituting different parts of DNA, both in genes and in junk DNA, can be thought of as *random* sequences of intricately correlated letters [9]. These correlations within the transcribed regions reflect the structure and function of proteins eventually produced from mRNA, while the intergenic space is organized under the influence of several competing factors: melting and bending properties of the DNA helix, genetic parasites such as self-splicing introns, neutral mutations, etc. where the full list of

[9]The level of correlations in the prokariotic DNA is comparable to that in the text of the "Hamlet". Eukaryotic DNA has a large component of uncorrelated white noise due to neutral mutations in the non-coding regions.

factors is unknown.

The spatial organization of DNA in chromosomes is essential for replication and transcription. For example, two segments which are far in the sequence but close in space (because of the folding) may be targeted by the same transcription factor.

Proteins. Proteins are *polymers* made of 20 basic amino-acids joined by *peptide bonds*. The length of natural proteins varies from a few dozens to several thousands amino-acids with a typical value of 200-300 amino-acids. An amino-acid is a compound consisting of two parts: a constant part formed by an amino group, a carboxyl group and a hydrogen atom (ACH), and a variable part, called *side chain*, which comes (in the existing organisms) in 20 flavors. Thus, a linear protein is a word in 20 letters represented by 20 amino-acids, where consecutive amino-acids are joined by peptide bonds at their ACH groups.

There is a formal similarity between polynucleotides and amino-acids. Both are *heteropolymers*, i.e. assemblages of several kinds of standard molecules, called *monomers*, having a connected chemically homogeneous (topologically linear) backbone, with chemically different short branches attached to each monomer of the backbone.

Different side chains have specific mutual interactions via weak bonds[10], which force a linear protein to *fold* in solution (under suitable temperature and acidity) into a particular geometric shape(s). There is a whole field dedicated to how and why proteins fold, with about 10,000 instances where the three-dimensional *conformation* (folding) is known (by a variety of bio-physical, bio-chemical and bio-informatical techniques). Yet, there is no clear conceptual picture on the nature, unicity and origin of the folding.

Roughly, as for RNA, the folding of a polypeptide chain comes as a solution of a system of stochastic differential equations where the final spatial conformation of the protein represents the minimum of the (free) energy of the molecule, incorporating weak bonding, bending and torsion energies,

[10] The essential weak bonds in folded proteins are hydrogen bonds, ionic bonds, van der Waals interactions and hydrophobic bonds. The latter is not a separate binding force but rather a result of the energy involved in the inserting of certain molecules into water. Weak bonds are sometimes reinforced by covalent S-S bonds between neighboring cysteine residues in a folded polypeptide chain. Cysteine (one of the two amino-acids containing sulphur) ensures the stability of snake poisons for example.

etc. A direct solution of this problem is computationally unfeasible and, in practice, one resorts to comparing a given protein to a similar one(s) where the structure is already known by using pattern matching/perturbation techniques. The true problem however, is not so much finding the conformation but identifying the *active sites*, that are specific locations on the external part of the three-dimensional conformation. An active site has a particular geometry and energy profile responsible for the enzymatic/binding activities of the protein, and involves one or several short peptide subchains of the protein. Most often, active sites are crevices of the protein accessible from the outside.

The "inverse folding problem", that is finding a word written in the 20 letters alphabet where the spatial conformation of the corresponding protein displays a site with a required property, lies at the core of the protein design, e.g. in pharmacology.

Proteins, carrying out the "program encoded by the genes" [11], run around the cell (or cell's compartments) and interact with other molecules: these can be other proteins and peptides (i.e. leftovers of partial degradation of proteins), (poly)nucleotides, lipids, polysaccharides and about 800 various kinds of small molecules. A large group of proteins called *enzymes* accelerate a variety of chemical reactions, other proteins make cell's architecture such as membranes and cytoskeleton, a third group regulates gene transcription, a fourth group transports molecules across the cell, yet another group is responsible for extracellular communication, etc.

Translation from mRNA to proteins. The formal aspect of the translation is quite easy: each amino-acid is coded by a triplet of bps called *codon*. As there are 64 such triplets, non surprisingly some amino-acids are coded by different codons. The translation begins with the recognition of a *start* codon, usually AUG, which determines the *reading frame*, i.e. a division of the mRNA chain modulo 3, and continues until a *stop* codon is found, which is usually either UGA, UAA, or UAG.

The codon–amino-acid correspondence is identical for most known organisms in agreement with the hypothesis that life on earth descends from

[11] This politically loaded widespread metaphor has the same kind of purposes and limitations as "the fittest survives" and ranks below "one gene $\longrightarrow$ one mRNA $\longrightarrow$ one protein $\longrightarrow$ one function".

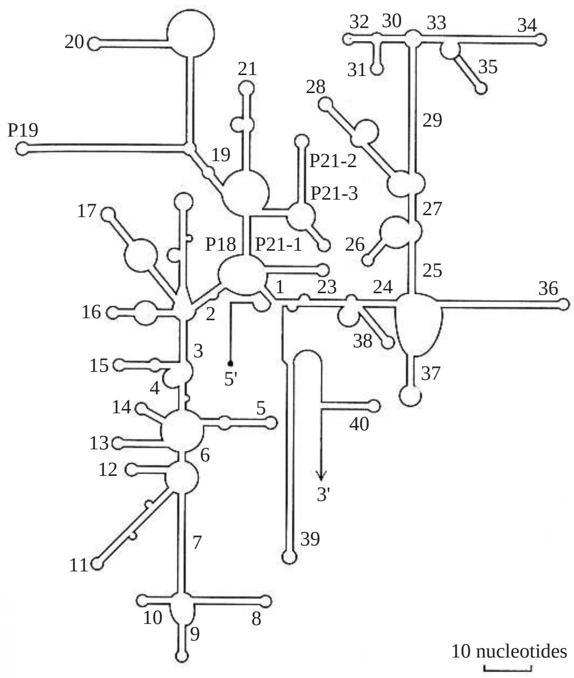

Figure 10: Folding (secondary structure) of ribosomal RNA.

a single macromolecular complex [12]. (Exceptions: many mitochondria and some protozoans differ in a few codons from the rest of organisms on earth. This is apparently due to the later evolutionary development.)

The actual translation from mRNA to proteins is a highly sophisticated mesochemical process (which can be reproduced in vitro albeit by far less efficiently than in the living cells). This is performed by two groups of macromolecular complexes, *ribosomes* and *transfer* RNA.

Ribosomes are roughly spherical, protein-synthesizing machines. They are huge by molecular standards, up to 30nm in diameter and about 3-10 times heavier than an average mRNA. They are composed of several different

[12]A molecule is a collection of atoms joined together by strong inter-atomic forces. A molecular complex is a conglomeration of several molecules kept together by weak (e.g. hydrogen bond) forces. With some stretch of imagination, a cell can be regarded as such a complex. Molecular complexes are regarded as dynamic rather than static entities, where the dynamics of the cell is structurally different from that of purely chemical complexes.

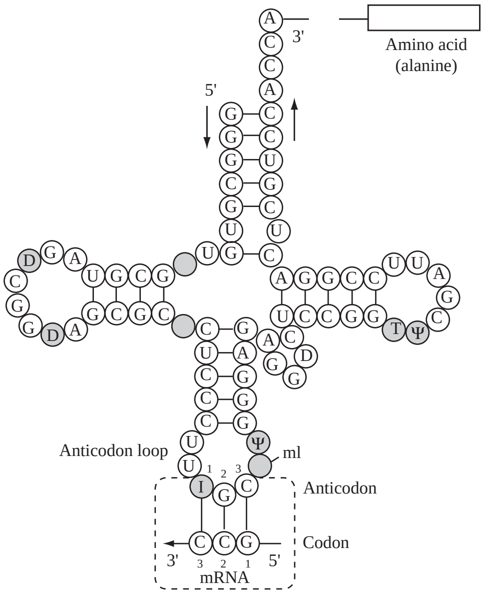

Figure 11: Folding (secondary structure) of transfer RNA of 76 nucleotides. The shaded letter I, for *Inocine*, in the anticodon, is a G lacking an aminogroup (i.e. $NH_2{-}C{\equiv}$ is replaced by $H{-}C{\equiv}$). This facilitates the non standard pairings C-I and A-I, and allows the cell to use fewer tRNA's than there are codons. DNA of all organisms contain 61 codons; bacteria have 30-40 tRNA's, and animal mitochondria manage with 22. On the other hand, animals and plants have 50-100.

ribosomal RNA molecules (rRNA) making about 2/3 of their weight, and of more than 50 proteins. Fig. 10 represents an example of rRNA.

A transfer RNA (for short tRNA) is $\approx$ 80 nucleotides long and folds in a particular way in obedience to the Crick-Watson complementarity [13]. It displays an *anti-codon* at a specific site somewhere in the middle of the strand and has an amino-acid attached to it (see Fig. 11). This complex is called *aminoacyl-tRNA*. The anti-codon is the triplet complementary to the one which encodes the amino-acid attached to this tRNA. The linkage between tRNAs and the corresponding amino-acids is performed by enzymes, called aaRS. There are exactly 20 of these (of size ranging from 500 to several thousands amino-acid residues) corresponding to the 20 amino-acids. Each aaRS brings together the tRNA and the amino-acid coded by the tRNA. The recognition of the appropriate tRNA by an aaRS depends on particular motives in the tRNA sequence, with a significant role played by the anti-codon. The pairing tRNA/amino-acid is the first major step of translation. The second step is the synthesis of the protein: very roughly, tRNAs get attached to an mRNA, the anticodon to the corresponding codon, thus aligning the amino-acids attached to these tRNAs. The latter are sewed together to a protein by a ribosome, at an average rate of 40 amino-acids per second for E. coli (and differs from organism to organism).

Metabolic pathways and genetic networks. A cell is a chemical machine in constant interaction with the environment. It takes from the outside nutrients [14] and oxygen (unless it is anaerobic) which are transformed into energy [15] that fuels the cellular machinery: transcription, translation and

[13]The spatial (tertiary) conformation of tRNA, and RNA in general, depends on more factors than sheer complementarity.

[14]The major source of energy for plant cells is not external chemicals but rather quanta of visible light which are transformed to chemical energy via photosynthesis. The sites of photosynthesis in plant cells and green algae are *chloroplasts*. These are independent entities, *organelles*, up to 10μm long and typically 0.5–2μm thick. In many respects, they are similar to mitochondria in animal cells. They both contain their own DNA and the proteins encoded by these are synthesized within their organelles. However most of the proteins in each organelles are encoded in nuclear DNA.

[15]Most of this energy is carried by small molecules, called *ATP*, consisting of adenine (the base A) joined with ribose (the sugar making the backbone of RNA), together with three phosphate groups. When ATP hydrolizes (i.e. reacts with water), it releases energy of $\approx 12 kcal/mol$ in the chemical environment of the cell. The hydrolises of ATP has a non-negligible activation energy and does not occur spontaneously without the presence

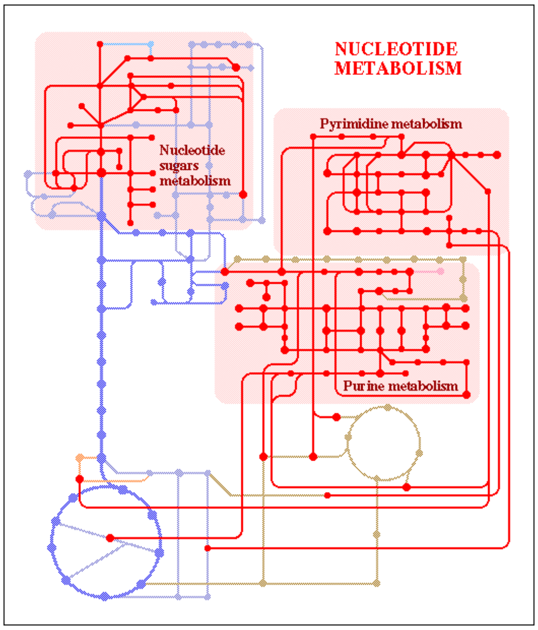

Figure 12: Schema of a small fragment of the metabolic pathway. (The direction of the reactions is not indicated.)

replication.

The processing of chemicals is organized into *metabolic pathways*. Schematically one might think of a directed graph $\mathcal{M}$ where the edges are marked by various chemical compounds, usually small molecules, and where vertices represent chemical transformations. (For an example, see Fig. 12) The direction of the flow in the graph indicates the predominant direction of the reactions reflecting the decrease of the free energy. Left to themselves, the chemicals in a cell would react extremely slowly. The rate of reaction are many fold times enhanced by high local concentrations within the walls of the cell architecture, and by catalytic effects of relevant enzymes present at the sites of the reactions [16].

Metabolic pathways are regulated by changing enzyme activity via *positive* and *negative* feedback loops. This allows the cell to sustain *homeostasis*, that is the (almost) constant level of the final product in the variable environment. In many situations, the first enzyme of a chain of reactions is inhibited by a negative feedback effect of the final product of the pathway. The inhibition is achieved by the binding of the final product to the enzyme, where the binding process is not ruled by ideal kinetics, but depends on the particular combinatorial arrangement of enzymes into complexes. This produces a highly non-linear dependence of binding on concentration with a pronounced threshold effect where a small excess of the final product may lead to an almost complete inhibition of the enzymatic activity of the protein. (This is called *collective allosteric transition* of proteins.)

---

of a catalyser. Enzymes serve as such catalyzers: they employ the energy of ATP to drive uphill reactions which need an import of free energy.

[16]The number of small molecules involved is quite large, their diffusion in cytoplasm is as fast as in water ($\approx 0.1 sec$ to go across a cell of $10 \mu m$), and their chemistry is governed, up to some extent, by the polylinear rules of the ideal kinetics: at a given temperature the percentage of chemicals being transformed to the final product is proportional to (certain powers of) their concentration; enzymes cannot change the direction of a chemical reaction but only the speed at which the process approaches equilibrium, i.e. the state where the rate of direct and inverse reactions mutually cancel each other. Besides, enzymes can promote energetically unfavorable reactions, for instance making a covalent bond between X and Y, by bringing ATP's (or other molecules carrying free energy like GTP) to the site of the reaction and by coupling the energy released by the ATP hydrolysis to the reaction between X and Y. Without an enzyme, such a process would be highly unlikely as it needs a meeting of three molecules X, Y and ATP, properly positioned with respect to each other.

The metabolic self-regulation process has a fast response time and is easily reversible. More radical regulation consists in changing the rate of production and degradation of enzymes, which is achieved by suppressing and/or activating the transcription of relevant genes. In fact, there is a second network, the *gene regulatory network*, responsible for this process, which is achieved with regulatory proteins binding to the regulatory regions of the relevant genes.

Roughly, a genetic network is a directed graph $\mathcal{G}$ whose nodes represent the genes. An arrow issued from a vertex is marked by the protein [17] coded by this gene, while incoming arrows tell us that the transcription of the gene is influenced by the protein. The current studies exhibit subgraphs of this graph with outcoming degree up to 100 and incoming degree up to 15. The most common motive in the network of E. coli is a simple negative feedback loop [18], by which a gene sustains a constant level of activity. More complicated patterns are being found as well.

A large amount of outcoming arrows is labelled by enzymes and ends up in the nodes of the graph of metabolic pathways. Some other arrows correspond to functional proteins involved into structure, transport, bio-chemical signalling, etc.

There is an extra combinatorial structure to the graph $\mathcal{M} \cup \mathcal{G}$ expressing the effects of small molecules on enzymes and regulatory proteins. The first example of the latter was discovered by Jacob and Monod in 1961, who found out that if the nutrient glucose is replaced by a somewhat less tasty lactose, the bacteria E. coli starts producing more enzymes needed for the assimilation of lactose (there are three of these enzymes). Normally these enzymes are produced by genes which are suppressed by a certain regulatory protein. The presence of significant amount of lactose in the cell disactivates this protein, and the lactose digesting enzymes are produced.

One can think of this auxiliary structure as a family of subgraphs in $\mathcal{G}$ parametrized by various bio-chemical and physical conditions of the cell, e.g. concentration of particular small molecules, temperature, acidity, etc. Each point in the parameter space distinguishes a relatively small subgraph indicating the part of the network active under a given condition.

[17] Our definition of a gene specifies an mRNA and thus the corresponding protein.

[18] Such a motive may turn out to be common for fast growing bacteria, but this is not the case for the fast growing eukaryote yeast, for instance.

The major mechanism of regulation consists in binding of proteins to regulatory regions which enhances or inhibits the transcription. We think of the binding as a (stochastic) mechanical process rather than a chemical one as it involves few molecules at a time. Yet, it is governed up to some extent by the rules of chemical kinetics and accidentally by inter-molecular quantum mechanical events [19].

The rough idea for an experimental identification of an edge in $\mathcal{G}$ marked by a protein P and pointing to a regulatory region R goes as follows: one replaces the coding part of the gene downstream of R (by means of recombinant techniques described later) with the code for a fluorescent protein with a short half life. When the gene producing P is activated, then the fluorescence seen in the cell indicates the effect of P on $\mathcal{G}$. Thus, one can detect enhancer and suppressor edges as well as oriented paths of such edges.

In principle, using microarrays (discussed later), one can detect some subgraph in $\mathcal{G}$ as well as evaluate the differences between such subgraphs sometimes without knowing each of them.

Post-transcriptional genetic regulation. The actual gene expression, that is the rate of production of proteins, is not only controlled in the course of transcription but also by the structure of RNA, by its location in the organism, and a variety of other mechanisms regulating translation. For example, a particular folding of RNA may slow down the translation process, the chemical environment may influence the rate of degradation of RNA, the usage of specific codons may effect the rate of translation by ribosomes. Apart from RNA, a cell may damp gene expression by degrading newly synthesized proteins.

Replication. The central event in replication is the production of two double stranded *daughters* DNA from a *mother* double stranded DNA: the two strands of the mother DNA separates and each of them serves as a template for a complementary strand, thus each daughter inherits one strand from the mother, with the other one being newly synthesized. The initial separation of the mother strands in a bacterium such as E. coli takes place with enzymatic help at a specific *initiation* site. The synthesis of the new strands

[19] A biologically relevant quantum event is the mismatching of complementary pairs leading to errors in replication. Also, some models of solvents (e.g. water peppered with small molecules and ions), relevant for binding of macromolecules, are based on semi-empirical quantum mechanical rules.

proceeds in two opposite directions starting from the initiation site, where the moving corners of the resulting growing "bubble" (made by the separated strands) are called *replication forks*. The synthesis of the new DNA should go along with the polarization of the template strands: only one of the strands of the mother DNA agrees with the direction of the movement of the fork, and templating the other strand is by necessity a discontinuous process which is divided into the synthesis of relatively short segments in the direction opposite to the movement of the fork. Systematic jumps of the double length of the segments allow the process to proceed in the direction of the fork movement.

The essential role in this process (we omit fine print) is played by DNA polymerase III, an enzyme built out of 10 subunits with a total mass $>$ 600kDa. DNA polymerase III proceeds with a rate of synthesis of ≈ 1000 nucleotides ($\approx 300nm$) per second in bacteria, and the full replication process, performed by two polymerase simultaneously (moving in opposite directions), takes about half an hour in E. coli with a circular DNA of $4,640$kb. The unlinking of the two daughter strands (associated to the two strands of the mother that are linked in space by the helical winding of DNA) is achieved with topoisomerase enzymes.

The synthesis of new DNA in human cells proceeds at ≈ 100 bases per second per fork. The full process takes about 8 hours with the number of forks estimated between 1,000 and 100,000, where not all forks are active simultaneously during the replication period. The newly born daughters move to opposite locations in the cell, the cell material is redistributed accordingly and the membrane undergoes topological modifications eventually leading to the division of the cell.

There are other replication-like processes modifying DNA besides doubling along with the cell division. For example, some segments may interchange their location within DNA, a (selfish) segment may generate several of its own copies within the same DNA, viruses may transport a segment from an organism to another, bacteria may exchange fragments of their DNA using plasmid vectors.

Macromolecular ensembles. Macromolecules in cells and sometimes in vitro are able to aggregate by self-assembly or by protein aided assembly into intricate geometric and rather rigid structures. Some are suspended in the cytoplasm such as *ribosomes*, *chaperones* and *proteasomes*. The latter are protein complexes of $\approx 2MDa$ serving in eukaryotic cells for selective degradation of proteins. Chaperones are protein complexes ($\approx 500kDa$) aiding protein folding by disrupting undesirable bonding between different polypeptide chains as well as between segments of the same chain.

The major global structures in cells are internal and external *membranes*, which are made of lipids, polysaccharides and proteins. They control the biochemical activity of the cell, sometimes actively participating in it, and also serve to separate different compartments of the cell. Also, they provide a support for several kinds of *molecular motors* such as H^+ driven "turbines" used in the production of ATP's and those serving for rotating the flagellae in some bacteria. Non-membrane based rotatory motors are apparently employed by some viruses (e.g. Bacteriophage $\Phi 29$) for packaging their DNA into a precursor capsid.

Taxonomic structures. The subject matter of taxonomy is constructing various metrics on the spaces of genotypes, species or other biological and biochemical entities (such as proteins and RNAs). There are two different, mathematically dual, approaches introducing a metric in the (moduli) space of structured objects. The first is the *evolutionary* (cladistic) approach, where the (phyletic) metric is defined by the length of the shortest path of elementary modifications of the object, similarly to the construction of intrinsic metrics by Gauss-Riemann-Kobayashi. Limiting such a metric to the space of actually existing *biological* entities often exhibits a pronounced tree-like behavior [20]. This is due to the huge size of the space of possible structures, e.g. of bp sequences of length 10^9, where the branches of the evolution process are very unlikely to come together and form cycles [21]. The second (phenetic) metric is the *phenomenological* one, similar in spirit to the Caratheodory metric on complex spaces. Here, we pick up some distinguished physiological parameters, e.g. size, style of breathing and nutrition, reproduction patterns, etc., and thus map our space into the space of parameters, where we choose some simple metric and induce it back to our original space. The dream of biologists is to independently construct these two metrics such that they would become equal on the space of the existing organisms.

[20] Every graph appears as a quotient of a tree, and the tree-likeness refers to the relative number of cycles created in the course of factorization.

[21] Yet, the tree structure is disrupted by the horizontal transfer of genes.

4 Scales and parameters

One wishes to describe a cell by first identifying and enumerating essential *structures* such as DNA, RNA, proteins, cytoskeleton, mitochondria, etc., presented in the previous section, and then specify *parameters* characteristic for a given cell or a particular state of a cell. When we speak about parameters, we think about points in a rather simple space. Typical parameters are real numbers taken from a given interval on the line. If we speak of two real parameters, we want them to be essentially independent and not subject to a complicated relation. Thus, we admit as a space of parameters a disc or a square in the plane but not a specific Cantor set like the Sierpinski gasket. However, if there is an evidence that the values of observable parameters are restricted to such an intricate set, then this set must be promoted from a parameter set to a new structure.

The number of relevant (types of) distinct structures in the molecular biology is rather limited, something of the order of $10^2 - 10^3$, but the dimension of the parameter space may be very large.

For instance, genomes in a first approximation are words in 4 letters: the structure is simple and transparent, the space of parameters for this structure is the space of 4 letter sequences of a certain length. The simplicity of the structure is due to the fact that we assumed that the individual parameters, valued at A, C, T, G, may take arbitrary independent values in all positions. The price we have to pay is the determination of *all* these letters for each individual DNA, with no help of additional structure limiting possible values of the parameters.

In the case of humans, the word has $\approx 3 \cdot 10^9$ letters: even reading ten letters per second (via chemical analysis) would take about 100 years. So biochemists learned to read real fast to (almost) determine the human genome in a mere decade!

There exists an extra structure in the space of (human) genomes: one chooses a prototypical sequence, a distinguished point in the space, such that any other sequence differs from it at relatively few essential [22] places. These variations between humans, probably constitute $\approx 10^4 - 10^5$ variable letters. Thus the genome of an individual human, modulo the "universal genome" is

[22] Here "essential" refers to the biological functioning of the organism. On the other hand criminologists distinguish different human genomes by variations of nucleotides in non-essential parts.

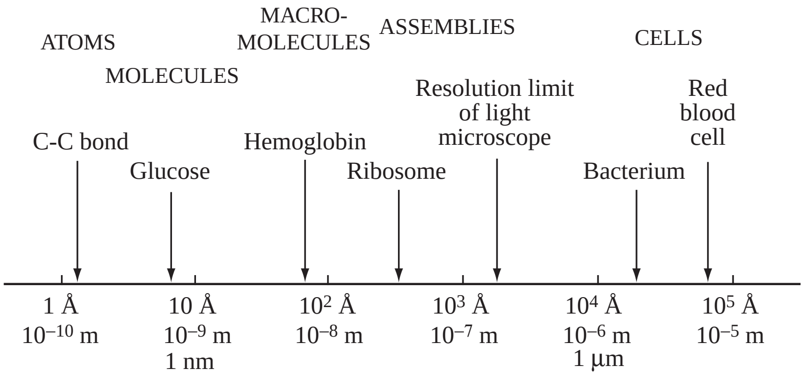

Figure 13: Table of dimensions.

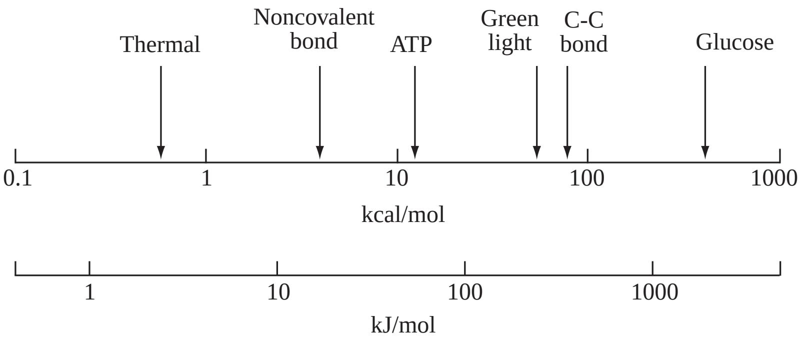

Figure 14: Table of energies.

essentially a random sequence of about fifty thousand uncorrelated letters.

Besides the above *coding* parameters, there is a variety of *physical* and *chemical* parameters characteristic for a functioning cell: size and mass of geometrically defined structures, characteristic time and energy of localized interactions, as well as temperature, concentrations of small molecules and ions, especially acidities (pH level), rates of reactions, etc.

The ranges of these parameters specify different structures of the cell, and the specific orders of magnitudes of the parameters apparently play a crucial role in the life of the cell. For example, there are three ranges of energy in the cell: thermal, weak (e.g. hydrogen bonding) and covalent, roughly of the order 0.6, 1–5 and 10–100 kcal/mol respectively. See Fig 13, Fig. 14 and

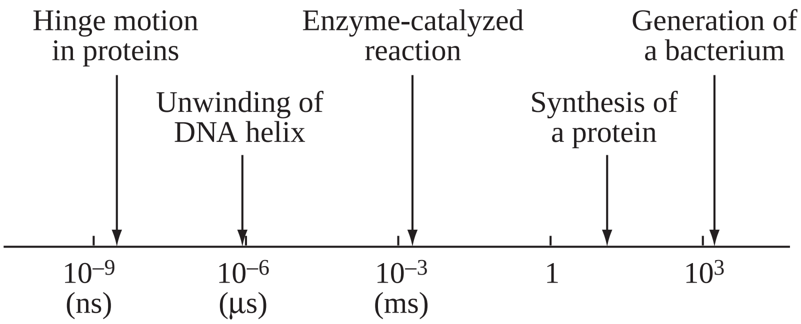

Figure 15: Time of processes (in seconds).

Fig. 15.

If we turn to more sophisticated patterns such as seen in genetic networks, the distinction between structures and parameters becomes less clear. The parameters specifying abstract networks are values 0,1 assigned to pairs of genes depending on whether they directly interact or not. However, this is not very useful as it leads to a huge parameter space of order 2^{10^9}. The basic problem is to isolate a structurally simple, admitting a reasonably short syntactic description, subset in the space of graphs (represented by 0–1 matrices) giving a sufficiently fine approximation to a realistic genetic network or a class of such networks. In reality, the structure of the network is augmented by additional combinatorial, numerical and functional ingredients: the space of parameters is enlarged, yet it helps to identify particular subclasses of networks actually present in cells. On the combinatorial side, we have a boolean structure indicating the enhancing/suppressing effect of a gene (or several genes) on another gene(s). This suggests clustering genes according to their role in regulation (e.g. position in the network, causality, general functions, co-regulation, etc.). Also, an individual network is immersed into a family of networks parametrized by the evolution tree. Browsing through the tree allows one (by working hard!) to identify persistent patterns corresponding to common functionality of networks across species. Furthermore, the feasible dynamics of the gene regulation (robustness, polylinearity of kinetics, etc.) imposes constraints on the network which may offset the extra complexity introduced by dynamic parameters.

The distinction between structures and parameters in biology is blurred compared to what happens in the physical sciences. A specific set of parameters on one "level of organization" may appear as a governing law at the "higher level". (An example in this spirit, is the emergence of symmetry in the growth of flowers according to Paul Green.) A possible mathematical formalization of that may appeal to dynamics associated to fractal potentials with sufficiently separated scales.

5 Design and control of macromolecules in vitro[23]

The bacterium E. coli, of volume $\approx 1\mu m^3$, is filled in by *cytoplasm*, a crowd of molecules in thermal motion. There are about 300,000 (non ribosomal) proteins of $\approx 10\mu m$ in diameter, 20,000 ribosomes [24] (25% of the cytoplasmic volume), 300,000 tRNA, a couple of thousands mRNA molecules, 50,000,000 small organic molecules including amino-acids, nucleotides, sugars, ATPs, etc. and various ions. The 70% of the volume is taken by $2 \cdot 10^{10}$ water molecules. (Eukaryotic cells, about 1,000 times bigger in volume, are filled in by a cytoplasm of similar composition, and they are architecturally organized by several relatively rigid structures: nucleus, cytoskeletons, Golgi apparatus, endoplasmic reticulum, vacuole, other organelles, etc.)

In the previous section we gave a rough sketch of what macromolecules do in the liquid crowd inside a cell, and now we turn to possibilities of what can be done biochemically in the test tube, or as biologists say, in *vitro*. Specifically, we want to explain the basic ideas behind bio-chemical manipulations which allow one to transform information encoded in polynucleotides into "visible" chemical and physical phenomena.

Imagine the (bio-chemically improbable) situation where we want to distinguish between two species of one-stranded polynucleotides. Both of them have the same number of bases, say one hundred, and they are represented in solution in two different tubes. Each tube contains a large ($\approx 10^{18}$) number of copies of the same molecule, and we want to decide which tube contains which sequence. The macroscopic properties of the two solutions are essentially indistinguishable since overall chemical and physical properties of two

[23]We limit ourselves to polynucleotides and do not touch upon proteins.

[24]The average gaps between ribosomes are $\approx$ their own size, and the gaps between (globular) proteins are $\approx$ twice their size. Proteins constantly hit each other (and run into ribosomes) with small molecules snicking in the gaps between them.

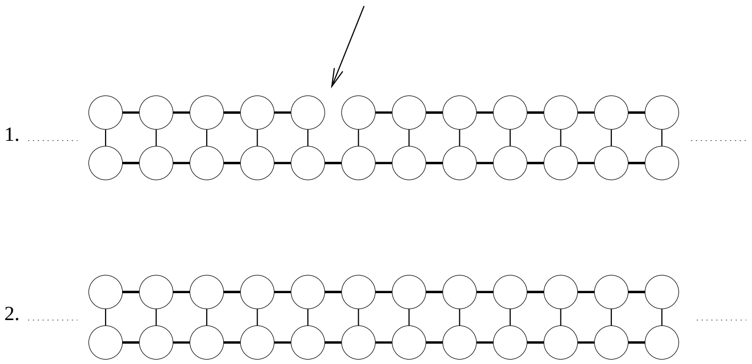

Figure 16: A nicked double stranded DNA (top), and the result of the action of ligase (bottom).

polynucleotides of equal length are very close to each other (at least if the sequences are not too special). However, if we recall the Crick-Watson duality, we come up with the following obvious solution: prepare a third tube with the double amount of polynucleotide molecules complementary to those in the first one, and pour half of its content in the first tube and half into the second. The complementary polynucleotides in the first tube will hybridize forming double stranded molecules, while in the second tube they will remain single stranded. Since the physical properties of the molecules in the two tubes became sharply distinct, this can easily be seen in the behavior of a variety of macroscopic parameters of the two solutions: viscosity, optical properties, specific heat capacity, resistance to degradation by single-strand specific nucleases, etc.

To go further, besides the Crick-Watson complementarity controlling the hydrogen bonds, we need to break and create covalent bonds. In the cell, this is done by a variety of enzymes. Some of them are "universal" and do not discriminate between specific bases, and some others are *site specific*, i.e. their action takes place where a particular short subword is present. Here are a few examples:

Ligase: given a double stranded DNA with a "missing" covalent bond in one of the strand, this enzyme "constructs" the bond. Besides nicks, it also repairs double stranded breaks, albeit less efficiently. See Fig. 16.

DNA polymerase: Consider a one-stranded DNA together with a short com-

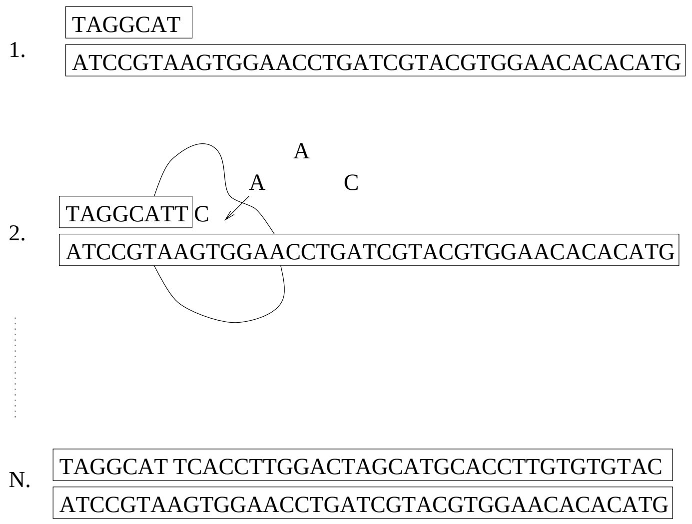

Figure 17: DNA Polymerase.

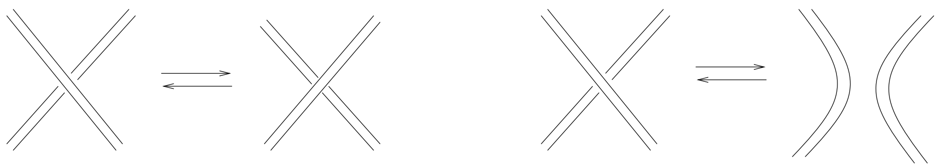

Figure 18: On the left, the action of topoisomerase II and on the right, the action of recombinase. The double arrow indicates that the operations are reversible.

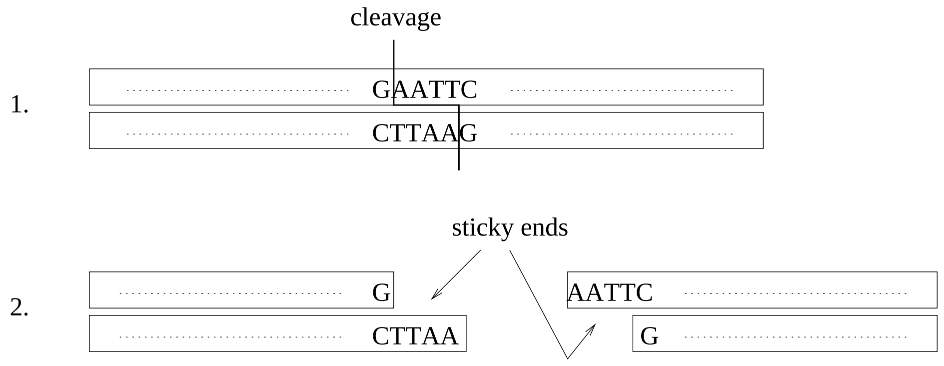

Figure 19: A double stranded DNA before and after the cleavage by EcoRI.

plementary word, called a *primer*, coupled to the "left" end of the DNA (for properly understood left/right on oriented DNA). Suppose that there is a supply of free nucleotides in the solution. Then, the DNA polymerase constructs the full complementary chain as illustrated in Fig. 17.

Topoisomerase II: when two segments of double stranded DNA come close together, the topoisomerase cuts the covalent bonds in one of the segments and then join them again on the other side of the second segment as shown in Fig. 18.

Recombinase: see Fig. 18.

Restriction enzyme EcoRI: it recognizes a double stranded DNA at the specific site GAATTC and cuts it as indicated in Fig. 19. Such a cleavage leaves free two short complementary single stranded segments, called *sticky ends*. Several cleaved DNA's can recombine by cross hybridization of the sticky ends. This step, followed by ligation, produces new DNA's.

DNAse I: it cuts double stranded DNA at sites specified by a certain class of subwords. The words of this class are characterized by *physical* properties of the corresponding stretches of DNA, especially by their flexibility. A precise determination of this class remains unknown, largely because of the complexity of the involved bio-mechanical problem.

The above enzymes can be used in vitro and allow the following manipulation with DNA strands.

Synthesis of polynucleotides. In order to chemically realize a given sequence, say TGCAATTCG, we start with T attached to a solid support by one of its ends and add a modified G to the solution so that G can covalently bind to the free end of T, but no G can bind to G. After TG is formed, the unattached G's are washed out from the solution, the "blunt" end of G is converted to its active form, modified C's are added to the solution, and so on. In order to have a large area of the solid support, one uses as a support microscopic beads suspended in the solution, or a microporous glass, with pores $\approx 1\mu$m and area $10^{18} nm^2$ per $1cm^3$ of volume, allowing in practice up to 10^{19} growing molecules on the surface.

One can build up oligonucleotides of 100-200 nucleotides in length. At every stage of the process, some oligo might not acquire the added base and the synthesis protocol includes a step, called *capping*, in which unreacted

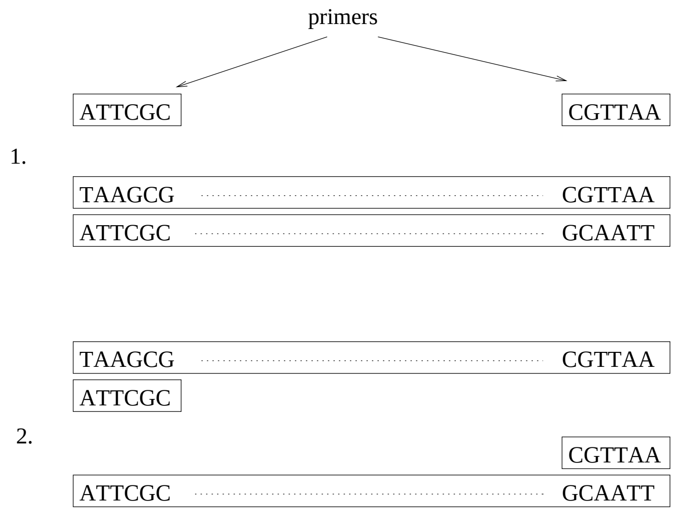

Figure 20: A double stranded DNA and the oligos corresponding to the beginning ends of its complementary strands. After the strands are separated, the oligos hybridize to the ends.

growing ends are "chemically disactivated" to prevent further growth. This prevents the appearance of long erroneous oligos (except for errors occurring at the end of the synthesis) and makes purification easier. Yet, the sequential accumulation of errors makes a reliable synthesis of longer oligos impossible by this process.

Polymerase Chain Reaction (PCR). This is in the league with the invention of the wheel and the nuclear chain reaction.

Given several molecules of a double stranded DNA in a tube, one can amplify their number *exponentially* in time along the following lines. Add to the tube N (of order 10^{19}) copies of two different single stranded oligonucleotides of the length about 10 bases, identical to the starting words of the two strands of DNA. Also add a generous amount of nucleotides and DNA polymerase. If we raise the temperature to $\approx 100°C$, the DNA denaturates, i.e. its strands fall apart. When we sufficiently lower the temperature, our oligos will hybridize to the corresponding complementary ends of DNA strands and will play the role of *primers* (this process will be in thermodynamical competition with the hybridization of the two strands among themselves, but since we have an excess of oligos, most DNA strands will hybridyze with primers). (See Fig. 20) Then the DNA polymerase synthesizes a complementary strand to each of our original strands and thus the total number of strands doubles. This works reliably even after many iterations for relatively short DNA's, several hundreds bps long, and there are various modifications of this method allowing reliable amplifications of DNA up to 30,000 bps long. The amount of DNA obtained is limited by the number of primers (and of free nucleotides, of course).

The presence of other DNA's, different from the ones beginning with a given word, represented by the oligos does not essentially affect the method: only the relevant DNAs will be amplified. This allows in particular to tell if a single molecule of a particular DNA is present in the tube among zillions of other DNAs. This is a useful test in biology as well as in criminology.

Sequencing. We start with a sample of one-stranded DNA (which can be amplified as needed using PCR) where we assume for simplicity of the exposition that this DNA is coupled with a primer at the beginning end. We add DNA polymerase, an excess of bases A, T, C, G into the tube and a small amount of A', that is a chemically modified A with the following property: if A' terminates a polynucleotide chain, then no new base can be attached to A' and the synthesis stops at A'. After polymerization, we obtain a variety of one-stranded DNA segments, complementary to initial subwords of the original DNA strands, where each segment terminates at A'. We separate the new strands from the old ones by raising the temperature and determine the spectrum of their masses or lengths by some physical/chemical method. For example, we can ionize our molecules (with reasonably controlled charges) and accelerate them in vacuum by an electric field, and follow their trajectories in a magnetic field. These are distributed according to mass and we can determine the masses of our segments with high precision. Alternatively, we can make the molecules going through a gel where their speed is inversely proportional to their length, and thus determine the length spectrum also with high precision. In either case, we determine the length of the segments terminating with A' which correspond to the positions of T's in the original DNA strand. This procedure applies to the remaining letters T, C, G and provides the position of all four bases in DNA.

Molecular beacons. Suppose we have a solution of various one stranded DNA (more realistically, RNA) in a tube and we want to decide whether a particular short word, say of 15 nucleotides, appears as a subword in some of them. One prepares a large amount of oligo of 25 bases which contains in the middle a complementary 15 letters word pinched between two mutually reverse complementary segments, each of 5 letters long. At room temperature, such molecules make *hairpins* with circular loops of 15 bases long and double stranded *stems* of 5 base pairs. (Compare to the picture in the right hand side of Fig. 7). If we add this oligo to the solution, each loop which finds the complementary word will hybridize with it, forming a rather rigid double stranded DNA. As a result, every such loop straightens up and the two ends of the oligo separate.

The separation of the ends is detected by chemically attaching a fluorophore to one end of the oligo and a *quencher* to the other end. When exposed to ultraviolet light, the fluorophore emits visible light unless it is close to the quencher which absorbs this light (and turns it into heat). Thus luminescence appears if the subword is present in DNA and some of the loops are open.

The oligo with the fluorophore and quencher attached to its ends serves as a *molecular beacon* signalling the presence of our word. This device is used, in particular, as a diagnostic tool for the detection of mutations in DNA (RNA).

Microarrays. These serve to determine the level of expression of N (possibly all) genes in a given cell culture or tissue by measuring the concentration of the corresponding mRNAs. Here is the idea. Take an array of N small tubes and put into the i-th tube a solution of molecular beacons complementary to a segment in the i-th mRNA. Add to each tube the solution of mRNA extracted from cells. Then, the intensity of light in the i-th tube will be (roughly, but not quite) proportional to the concentration of the i-th mRNA in the solution. In practice, one does not use molecular beacons (yet) but rather attaches fluorophore to the mRNA extracted from the culture, and measures fluorescence resulting from hybridization by different means (without the use of quenchers). To increase reliability and compensate for errors (partly arising from the drive for micro-miniaturization), one often compares expressions from two different cultures, marked by different fluorophores and mixed together. This mixture is added in each small tube and the color of the fluorescence witnesses the ratio of the expression.

For basic research, the main purpose of microarrays is to reconstruct the gene networks in a cell, but this goal faces many unsolved problems. The present microarrays are not sufficiently reliable, due to a variety of uncontrolled errors: self-hybridization of RNA affecting the hybridization with the probe, dependence of hybridization on the fluorescence dye, difficulty in synchronizing cell cycles and/or making time dependent measurements (the latter problem does not exist for molecular beacons), presence of dust particles and/or spontaneous fluorescence, etc. The major problem is that, due to post-transcriptional regulation, the production of a particular protein is not proportional to that of the corresponding mRNA and thus mRNA microarrays should be eventually complemented by protein-chips, where a protein is detected by binding to its antibodies, i.e. a specifically designed protein to which a given protein binds. Even granted that all these technical problems are resolved, the full combinatorial problem of reconstructing the gene network from the averaged (!) expression levels, under variable conditions, remains wide open [25].

Recombinant techniques. One can use the bacterial replication machinery for cloning *long* strands of DNA, which allows a higher fidelity than PCR. This machinery can be also used for generating and selecting sequences coding for proteins with desired enzymatic and/or binding properties, e.g. antibodies for specific pathogens.

Given a segment of double stranded DNA, up to several tens of thousands base pairs long, it can be inserted into the genome of a (bacterial) cell by a variety of recombinant techniques, also called *genetic engineering*. This is done with either *plasmids*[26] or *viruses*[27] used as *vectors* carrying the desired DNA segment into the cells. The vector DNA is cleaved with a restriction enzyme and ligated to the DNA segment. (The latter is prepared with com-

[25]These difficulties do not play a significant role in certain practical applications of microarrays such as fast genotyping, where direct measurement might be sufficient, and in some diagnostic procedures, where a rough gene clustering suffices.

[26]Plasmids are circular double stranded DNA molecules disjoint from the chromosomal DNA in cells, and vary from a few thousands to more than 100 thousands bps in length. They use the translation machinery of the (host) cell and replicate along with the cell division. They occur naturally in bacteria and yeast, for example.

[27]A virus particle, called a *virion*, is a self-assembling macromolecular complex constituted of one or several polynucleotide molecules covered by a protein coat. Viruses penetrate into prokaryotic or eukaryotic cells, throw away their coats and use the cell machinery to replicate and to produce viral proteins (the number of proteins encoded in a virus ranges from as few as 4 up to roughly 200). Some viruses live in the cytoplasm and some incorporate their DNA into the chromosomal DNA of the host.

plementary sticky ends which allow the hybridization of the segment to the vector and the following ligation.) Then, the resulting *recombinant vector* is inserted into the cell. This is done either with the help of a chemical making the membrane permeable to plasmids, or by shooting particles coated with plasmids into the cell culture. A viral vector penetrates into the cells by itself using its coating proteins.

A common technique for selecting the cells which did receive recombinant vectors, consists in inserting an auxiliary gene into plasmids, where the gene makes the receiving bacterium resistant to a particular antibiotic. When the treated cells are raised up on a nutrient containing the antibiotic, only the transformed cells will survive. In protein engineering, one "marks" transformed viruses by inserting the code of a specific protein next to the gene of a coating protein. In the course of translation, the new protein will *fuse* with the coating protein and will be displayed on the coat of the virus. Then one prepares a "complementary" protein with a special affinity to the fused protein and attaches it to a solid support. When the support is brought to contact with the viral culture, the marked viruses bind by affinity and are extracted from the solution.

Apologies. The above description of biochemical techniques as well as our overview of the cell functions should not leave an expression of being anything close to exhaustive or in the front-line of research. There is a body of classical techniques such as NMR, x-ray cristallography, relaxation spectrography, emission spectroscopy, and various microscopy techniques: electron, cryo-electron, scanning transmission electron, scanning tunneling, atomic force, etc. We did not touch upon proteomics, microscopy for subcellular localization of fluorescent markers, design for unicellular studies, single molecules experiments, and many many other directions.

6 Formalizable structures

One seeks for a *coherent* [28] network of *models*, i.e. consistent mathematical theories, reflecting the behavior of macromolecules and of ensembles of those in vivo and in vitro.

[28] The inter-connections between theories do not have to be fully formalized. Sometimes, one does not even expect that they are formalizable at all, as for the "quantum $\longleftrightarrow$ classical" correspondence, for instance.

Examples of models in physics are classical thermodynamics, equilibrium statistical mechanics, the spectral theory of Schrödinger equations, etc. Here one is often keen at deducing simple macroscopic (e.g. thermodynamical) laws from given local microscopic interactions. In contrast, a biologist seeks for an explanation of the complexity of the observables of macromolecular systems in terms of simple local laws. Besides, he tries to isolate regular macroscopic and mesoscopic patterns in the behavior of macromolecular and, in particular, biological systems in order to predict and design experiments.

Let us start with a list of patterns which invite a mathematical framework:

Repeativeness and partial symmetry of stochastic motives. The very existence of biology (and any rational science in general) depends on the systematic repetition of patterns, motifs and structures, which allows *compression* of the observed information. Given a large collection of biological macromolecules, macromolecular complexes, cells, organisms etc., one can divide them into relatively few groups with great similarities between members of the same groups. In fact, this phenomenon starts from biochemistry, where the number of small molecules in a cell by far exceeds the number of different species of molecules (i.e. amino-acids, nucleotides, sugars, etc.), and where large molecules (i.e. polynucleotides, proteins and polysaccharides) display similar repetitiveness of basic motives. Besides, even literally different bio-molecular aggregates, e.g. cells, share many common features in their design and functionality. This suggests that the same low variability of combinatorial and dynamical patterns will emerge in other structures, e.g. gene networks, once these are revealed.

The symmetry in biology is of a different nature than that in physics, where one sees statistically completely homogeneous media such as gases and liquids as well as truly homogeneous symmetric structures in crystals. In biology, *stochastic* patterns, e.g. organisms of a given species, appear faithful to fine details by far more often than one might naively expect on the basis of physics and probability theory. Yet, this phenomenon should be eventually explained in terms of locality and stochasticity. More specifically, why accidental low free energy metastable states harbor stochastic symmetry? How the major biological sources of symmetry, which are *templating*, *replication*, and *universality of production*, can be derived from general energy/entropy considerations?

Channeled relaxation. In many situations a macromolecular system minimizes a complicated function of its parameters in an amazingly short time: RNA and proteins fold in several minutes [29], macromolecular complexes self-assemble in cells, and sometimes in vitro, almost as fast, covalent bonds are destroyed and created in a fraction of a second in the presence of enzymes, the reproductive efficiency of bacteria has grown from 0 to 2 divisions per hour in mere two billion years! Eventually one wants to understand this phenomenon as a feature of the function being minimized, or rather of families of functions.

Complementarity and self-assembly. Both transcription and DNA replication are based on templating by the Crick-Watson complementarity. Also, binding of proteins (e.g. between a protein and its antibody) can be seen as a complementarity pairing between macromolecules.

Mathematically speaking, there are (partial) involutive symmetries in the macromolecular world, which bring to one's mind reflection groups. For example, the coats of viruses with icosahedral symmetry, self-assemble this way, out of several identical protein molecules [30]. Less symmetrically, ribosomes self-assemble, even in vitro, if the RNAs and proteins constituting them are present in the solution.

It happens sometimes, e.g. for viruses, that the process of self-assembly goes in stages, where the initial assembled structures serve as a scaffolding for the following one and then disappear from the final result.

A simple model for self-assembly is provided by a collection of convex

[29] This should be confronted with the rate of individual molecular events: the hinge motion of proteins happens on the scale of nanoseconds while molecular vibrations are thousands, sometimes millions, time faster. However, these speeds cannot significantly shorten the rate of folding of long polypeptides since the number of possible configurations grows exponentially with the length: to explore 2^N configurations of chains of length N with a rate of 10^{15} moves per sec, one needs 2^{N-50} seconds. On the other hand, the number of deep and wide local minima may grow significantly slower than 2^N, thus allowing fast folding by the gradient directed random walk. An argument due to Eigen (private communication) and Kauffman suggests that adding new dimensions to the configuration space, predominantly turns local minima (of the energy function) to saddle points and thus a random function in many variables may have fewer "dangerous" local minima than a naïve counting suggests. A comprehensive mathematical development of this idea is still missing.

[30] Chirality of biomolecules does not allow (orientation reversing) reflections, but rotation groups do appear as symmetries of crystallized proteins and of coats of some viruses, for example.

(flexible) polyhedra in the space, with prescribed affinity between some of their faces. Thrown into solution, these polyhedra may (or may not) assemble into larger structures similar to covering of orbihedra.

Compartmentalisation. A cell, especially a eukaryotic one, is far from being chemically homogeneous. It is divided into semi-connected compartments and it is filled with filaments channelling and enhancing the biochemistry of the cell. Both the walls of the compartments and the filament channels can be *static*, kept together by chemical bonds, or *virtual*, supported by the dynamics of the cell. Compartmentalisation inhibits the global relaxation of the system but enhances partial relaxations within the compartments. The cell itself is defined as a compartment formed by a semi-permeable membrane, decoupling the dynamics within and without the cell.

One wishes to see this picture in a high dimensional (time-)space as a kind of *mediating topology* depending on auxiliary parameters such as the permeability of a given substrate through the membrane.

Homeostasis and replicative stability. Homeostasis, a chemio-dynamical stability of a cell in the variable environment, is disrupted by replication. The cell is brought out of a stable regime, dynamically speaking an *attractor*, but the stability is regained in a different framework: the dynamical features of the system reappear in the *multiplicity* of nearly identical cells represented by the cartesian product of many copies of the attractor, and ensure preservation of information by perpetuation. In fact, homeostasis of an individual cell cannot be stable for a long time [31] as it would be destroyed by random fluctuations within and without the cell. There is no adequate mathematical formalism to express the intuitively clear idea of replicative stability of dynamical systems. (A possible model might use dynamical time of variable fractal dimension depending on the number of bacteria in the population, expressing the idea of weak correlations between time-clocks of different bacteria.) It is unclear if the appearance of many copies of similar individuals is unavoidable or just a transitory feature of life. The question is whether there exists, mathematically speaking, a stable highly organized system with no high repetition of structural components, similar in spirit to *Oceanus sapientissimus* of Stanislav Lem.

[31] In a protected environment, a cell, such as a neuron in the human brain, may live for a 100 years.

Information and amplification. Among other biochemical parameters of the cell, the information encoded in DNA is distinguished by the following features:

- it is time conserved [32]. More precisely, it changes with much slower rate than other parameters;

- unlike most quantities conserved by physical systems (e.g. energy, momentum, etc.), the "information observable" ranges in a very large space: the space of 4 letters sequences of significant length;

- small perturbations of information unpredictably amplify. The dynamics of the cell maps small "information sets" onto much larger spaces of biochemical and morphological parameters. Besides expanding the size, this map may dramatically increase the structural complexity of the information set.

What is a proper mathematical description(s) of the paradigm "genotype determines phenotype"? Would every consistent mathematical model require a notion of "inheritable initial condition" corresponding to what biologists call "epigenotype"? Can the above properties be turned into precise definitions? If there is a rigorous model, are the three properties mutually independent in it? Is the presence of information/memory unavoidable in any life-like dynamics?

Universality of production: translation and replication. The two basic cell machineries, DNA replication and synthesis of proteins, are reminiscent of the universal Turing machine. DNA replication works on any given piece of DNA (viruses know it only too well). The same almost applies to transcription and translation: DNA is roughly divided in two parts. One encodes *house-keeping genes* coding for proteins supporting the basic machinery, e.g. ribosomal proteins. This house-keeping DNA cannot be arbitrarily modified without badly disrupting the cell functions. On the contrary, one can change the remaining part of DNA by inserting and deleting DNA fragments, thus making the cell to synthesize any given protein (unless it happens to be poisonous to the cell).

Can this universality and interchangeability be expressed in general mathematical terms, e.g. in the language of dynamical systems as a pseudo-group (or something like a category) of partial symmetries?

[32] An essential property of information which is commonly emphasized, is its *syntactic invariance*, that is the relative independence of the specific physical carrier. Such a property can be hardly formalized in a dynamical framework.

Separation of energy scales. The functioning of the cell can be thought of as a continuous flow and redistribution of energy which comes in three essentially different forms: *thermal* energy, (weak) *binding* energy, and *covalent* energy.

The *thermal* energy is a random field in space, where (not very) large fluctuations are comparable to the binding energy but much weaker than the potential barriers encompassing the covalent energy. If not for the mediating effects of binding energies, there would be no significant thermal dissipation of the covalent energy (at the room temperature) on the time-scale of the cell cycle [33]. The system would behave as if it were at equilibrium. A familiar example is a water solution of H_2O_2 which is rather stable at the room temperature and away from light; enzymes from saliva act as catalysers and make H_2O_2 to decay with a release of free oxygen visible in the bubbles if one spits into the solution.

For us, the *covalent* energy is a *scalar* distribution of labelled points in space, where the geometric polarization of the bonds is ignored and where each label corresponds to a species of small molecules. The label should specify the covalent energy of a molecule as well as the *activation energy*, that is the height of the potential wall preventing the release of the stored energy [34]. The relevant information is encoded in density functions of labeled points in the 3-space.

The *binding* energy applies to interaction between macromolecules and other large and small molecules. (The weak interactions between small molecules are ignored here, or delegated to the thermal energy.) Unlike the thermal and the scalar covalent energy, the binding energy is polarized and depends on the mutual position of the interacting molecules.

The system governed *only* by thermal and binding energies (without gain or loss of covalent energy) would relax to the thermal (quasi-)equilibrium relatively fast compared to the cell cycle [35]. Deadly on the scale of the cell, the "rigor mortis" following relaxation, is incorporated into the local dynamics for assembling skeletal structures with specific geometries depending on the

[33] Some covalent bonds can be spontaneously broken by hydrolysis with a release of thermal energy, but we do not consider this at the moment.

[34] This is an oversimplification: some covalent energy can be stored in macromolecules. Besides, it may "reside" in *pairs* of molecules as in the gaseous mixture of H_2 and O_2 for example.

[35] Death by starvation and asphyxiation maybe postponed by sporulation.

polarization of binding energies.

The structure produced by the binding energy channels the distribution of the covalent energy and facilitates its exchange (metabolic pathways) and release. The released covalent energy is harnessed to build up macromolecules. Some part of the covalent energy is redistributed among (macro)molecules, some is converted to the binding energies of the macromolecules, and the rest dissipates to heat. The changes in binding energies bring in dynamics to the geometric structures created by them.

It seems that no organized structure can exist in the presence of only two levels of energy. The sizes of the gaps seem also important: it is hard to imagine self-organizing structures employing nuclear energies [36] instead of covalent ones, but there is no mathematical model where this would become a theorem.

Specificity of scales and numbers. To demonstrate the importance of specific relative values of numerical parameters in the functioning cell (numbers of particles, their size, mass, energies and interaction/relaxation time scales) we shall draw a conclusion from the following simple observation: there are as many microns (a bacterium size) in a centimeter (a small tube size), as angstroms (an atom size) in a micron.

"Theorem": *Consider a* random *pond of water of* $1000m^3$ *containing the chemical compounds needed for life and an excess of free energy. Then, with probability* $P \geq 1 - 10^{-10^2}$*, it contains life in unicellular form.*

The proof depends on a "lemma" and an extra assumption.

"Lemma": *There exists a bacterium* $B \leq 1\mu m^3$ *in volume where the probability of replication within a unit time interval* Δ_t *is 10 times greater than the probability of death within* Δ_t.

To rigorously prove the lemma one needs a *formal* description of a bac-

[36] The sources of nuclear energy, external to biological systems, such as the sun and the inside of the earth, are crucial for sustaining life. These would be deadly in the pure form of high energy photons and they become acceptable only after intermediate transformation to the thermal energy corresponding to green light photons and below. The problem is whether the nuclear energy can be structurally incorporated into a life-like system where a high level of energy is tempered by sparse spatial distribution but without transformation to lower level thermal energy. The obvious intuitive answer is negative and it would be nice to have a general mathematical theorem confirming the intuition.

terium B, e.g. E. coli. We do know that E. coli exists [37] and the biologists are currently preoccupied with furnishing such a description.

"Simplifying assumption": *When a bacterium divides, its daughters are separated in the pond and do not interact with each other except for sharing common nutrients. In particular, one bacterium cannot kill another one, even allowing mutation and evolution.*

This assumption does not correspond to the evolution of unicellular organisms on earth: one of the multicellular descendents of the primordial cell is nearly able to wipe out the bacterial population inhabiting the $10^{18}m^3$ "pond" of the biosphere.

"Proof": Since $1\mu m^3$ contains $\leq 10^{12}$ atoms (and small molecules such as H_2O, CO_2, N_2 and with luck NH_3), the probability of a spontaneous assembly of B out of atoms, within Δ_t, e.g. one hour, is $\gtrsim 10^{-12\cdot 10^{12}}$, since each atom out of 10^{12} can occupy 10^{12} positions in space. On the other hand, the pond can be happily occupied by 10^{16} bacteria B, where each $1\mu m^3$ bacterium is allowed $10^5\mu m^3$ living space.

Since $1/2 \gg 1/10$, once B appears, its descendents populate the pond with probability $P \geq 1/2$. Also, the probability of death of the full population of bacteria in the pond is $\leq 10^{-10^{16}}$. Hence, $P_{life}/P_{death} \geq 1/2 \cdot 10^{10^{16}} \cdot 10^{-12\cdot 10^{12}}$. Q.E.D.

Remarks. If we scale the size of organisms linearly by k, the required living space grows as k^2. For example, to make the above work for a $1cm$ size organism, one would need a volume of $\approx 10^{20}km^3$ which by far exceeds the volume of earth.

The above "theorem" does not say that life appears from non-life with high probability in a short stretch of time but rather that the average number of "alive states" (i.e. states where the system contains a living cell) in the configuration space exceeds that of the "dead states". The paradox appears

[37] E. coli bacteria have volume $0.6 - 3\mu m^3$ and they are not able to sustain their existence without the support of photosynthesizing bacteria. The latter, the present day cyanobacteria, are rather large, $\approx 100\mu m^3$. On the other hand, there are other bacteria, called *mycoplasma*, that can be as small as $0.02\mu m^3$. But these usually live parasitic existence in association with animal and plant cells. Cells of the size of mycoplasma and with metabolic abilities of cyanobacteria are ideal for our lemma. Such cells are believed to be among the first inhabitants on earth.

only if one misuses the ergodic theorem and interprets the time average as something observed in realistic time intervals. In our example, huge time intervals where the system has no life will interchange with even larger intervals with life. The boundary of a long interval is negligibly small compared to the length of the interval, making the transition probability very low. But since the action takes place in a high dimensional configuration space, the boundaries of relevant regions (i.e. regions of alive states) become much larger relatively to the volume of these regions (isoperimetric concentration phenomenon). Intuitively, the large dimension allows many scenarios for non-life/life transition. See the end of Section 7 for a mathematical discussion.

The probability of a configuration of atoms depends on the energy via the Gibbs factor. This has no significant effect on our computation but the presence of potential barriers (activation energies) aggravates the transition problem from "dead" to "alive" states.

The probability of an "alive" configuration is by far greater than 10^{-N} for $N = 10^{13}$, since there is more than one "alive" configuration: the potential number of possible self-replicating configurations is something like $10^{\alpha \cdot N}$, where α is a small, but not negligibly small, positive number. This α roughly represents the percentage of rearrangements of atoms keeping an "alive" system alive: out of N atoms, $\alpha \cdot N$ of them can be independently modified.

The number α can be interpreted as a *relative dimension* of "life" and its realistic evaluation would be quite interesting. The idea of dimension can be seen in a $M \times M$ regular square array of points in the plane, where $M = 10^{\frac{1}{2}N}$. Subsets containing $10^{\alpha \cdot N}$ points for $\alpha = \frac{1}{2}$ corresponds to lines in the square and have relative dimension $\frac{1}{2}$.

Self-replication can be thought of as a fixed point of a certain transformation (dynamical system) modelling replication. A certain set of replicative systems make an attraction basin to the set of (stably) self-replicative systems, and the relative measure of this basin maybe reasonably large to allow a time realistic scenario for transition from non-life to self-replicating life. Replication may represent, for example, a division of a membrane-bounded entity (e.g. a micelle or a liposome) into daughters far from being identical to the mother and to each other, and in the course of sequential divisions, they structurally converge to self-replicating organisms.

All ten features mutually intertwine and our ... of them will not be linear.

7 Models and problems

Populations of strands and hybridization diagrams.

An *AG-strand*, or just a *strand*, is a directed finite linear graph with letters A, A^{-1}, G, G^{-1} labelling each edge. This is either an oriented segment or a cycle (a topological circle) with letters written on it. In other words, it is either a path or a cycle in the figure $'\infty'$ representing the free group on two generators.

A *population* of strands is a measure on the set of strands. *Warning*: as we want to model populations of strands in realistic solutions in tubes, where the measure weight attached to a strand equals the number of the copies of the strand in the solution, one must take into account the fact that the number of strands in a population $(10^{20} - 10^{30})$ is large compared to the number of all possible short strands $(30 - 40$ letters long) but abysmally small compared to the number of strands of length more than 100 letters. A better model for a population of long strands is a *random measure* on the space of strands.

A *hybridization diagram* D is a graph $\underline{D}$ together with a combinatorial map of a disjoint union of strands onto it, denoted $h : \bigsqcup_{i\in I} S_i \to \underline{D}$, where the S_i are strands indexed by I, and where the following conditions are satisfied: vertices go to vertices and edges *onto* edges, thus every strand S_i is realized by a path or by a cycle in D. Each edge in D has as a pullback either a single edge or two edges with opposite orientation and reciprocal labels. The latter is called *double stranded edge* and the former, is called *single stranded.* [38]

Two edges in D are *strand connected* if they both descend from the same strand S_i by the map h. A hybridization diagram is *connected* if every edge can be reached from another one by a chain of mutually strand connected segments. (See Fig. 26.) Thus, every diagram divides into connected components which can be a priori smaller than the topological components of D. We specifically assume that the underlying topology of D actually defines the same components as the strand connectedness.

Hybridization (involution). Two edges in the union $\bigsqcup_{i\in I} S_i$ are *hybridized* if they have equal images under h in D. Hybridized edges come in reciprocal

[38] We do not distinguish D and $\underline{D}$ insofar as it does not lead to confusion.

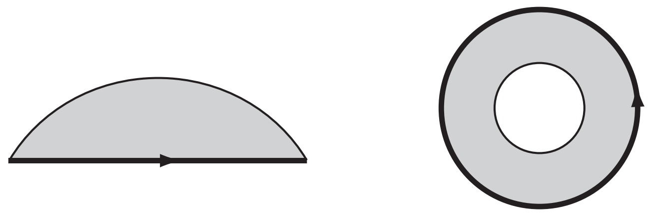

Figure 21: Surfaces associated to a strand (left) and to a circular strand (right).

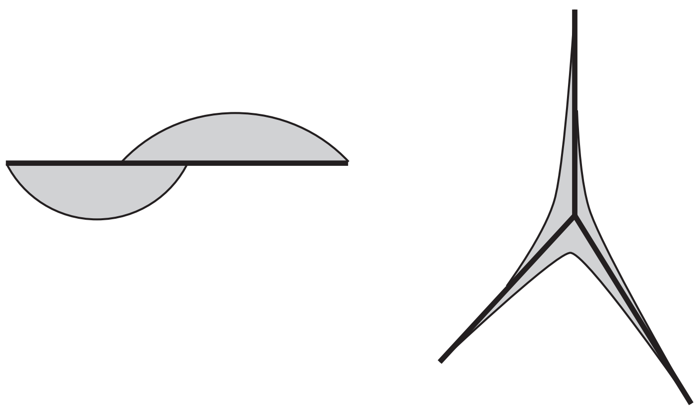

Figure 22: Two examples of hybridization surfaces.

pairs with a natural involution $\tilde{h}$ on the union $S_{hy} \subset \bigsqcup_{i\in I} S_i$ of those pairs. The graph D can be described in these terms as the quotient of $\bigsqcup_{i\in I} S_i$ by the involution $\tilde{h}$.

Hybridization surfaces. Given a strand S, we consider the surface C_S, that is defined as $S \times [0,1]$ for circular strands, and for linear strands as $S \times [0,1]$ with the two edges corresponding to the end points of S shrunk to single points. (See Fig. 21.)

The *hybridization surface* C_D *associated* to the diagram $D = (\underline{D}, h : \bigsqcup_{i\in I} S_i \to \underline{D})$, for $S_i = S_i \times 1$, is defined by attaching the disjoint union of the surfaces C_{S_i} to D via the map h. This is indeed a topological surface containing $\underline{D}$ and naturally retracting to it. The boundary of C_D is subdivided in two parts: one corresponds to the non-hybridized part of the strands S_i's, and the second part, i.e. $\bigsqcup_{i\in I} S_i \times 0$, decomposes into segments with disjoint interiors corresponding to the $S_i \times 0$. (See Fig. 22.) The labelling map $\bigsqcup_{i\in I} S_i \to {}'\infty'$ naturally extends to a continuous map $C_D \to {}'\infty'$. Conversely, a hybridization diagram can be *defined* as a surface with a distinguished oriented part of the boundary and a labelling map of this surface to $'\infty'$.

The *population of diagrams* is a measure on the set of connected hybridization diagrams. It represents the numbers of species (i.e. isomorphism classes) of macromolecular hybridization aggregates in a solution. Here, even more than for strand populations, one must be careful with the stochastic interpretation of this measure. A true stochastic object is a *random population* of diagrams, where the weights assigned to the diagrams are not real numbers but positive random variables, or better, it is a probability measure on the space of populations of diagrams.

Dynamics of hybridization. Dynamics of hybridization refers to a random walk in the space of diagrams, reflecting what happens to an ensemble of strands in a solution. At every step of the walk, a diagram transforms to another diagram with a certain probability weight assigned to the step, where each transformation is a hybridization or a dissociation of an edge. A full description of such a walk should take into account the interaction energy between different letters, corresponding to actual energies of hydrogen bonds, the bending energy, etc.

The stationary states of the corresponding Markov chains can be modeled by a Gibbs measure (defined later in the section), with proper entropic

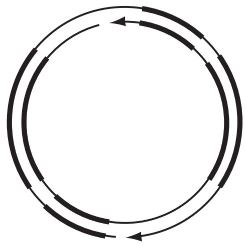

Figure 23: Circular hybridization of two non-circular strands containing non-trivial symmetry in their words.

weights assigned to these states. The first question (corresponding to the zero-energy) is finding the minimum energy diagrams. Let us simplify further, ignore the bending energy and prescribe the (local) energy at every *edge* as follows: every non-hybridized edge has energy zero; to every hybridized AA^{-1} pair is assigned a certain negative energy $e(A)$, and to every hybridized GG^{-1} pair is assigned a negative energy $e(G)$. Besides, one can partly take into account the bending by assigning certain positive energies to the *vertices* of the graph. For example, if the hybridization map folds (i.e. it is not locally one-to-one) over some vertex in D, then one assigns infinite positive energy to this vertex to exclude such a folding.

The total *combinatorial energy* of the diagram is the sum of energies of all edges and vertices. In some cases, the minimum energy diagram can be easily found. Here are some examples.

Complementary strands. If S and S^{-1} are complementary strands, then obviously the minimum of the energy is achieved by the linear or the circular diagram hybridizing S with S^{-1}. This minimal energy diagram is *unique* if we agree to assign (arbitrarily small) *positive* energies to the vertices, except for circular strands where the corresponding word has a non-trivial cyclic symmetry, and where the number of non-isomorphic diagrams equals the order of the (cyclic) symmetry group of our word with the two ends identified. (See Fig. 23.)

Triple junctions. Consider three sufficiently long random words X, Y, Z, and take the three strands XY, $Y^{-1}Z$ and $Z^{-1}X^{-1}$. The minimal energy diagram is unique and it is depicted in Fig. 24.

These two examples indicate how the presence of symmetry and of ran-

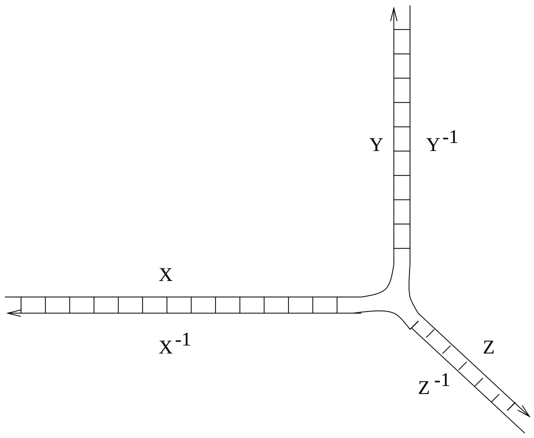

Figure 24: Minimal energy hybridization of the three words XY, $Z^{-1}X^{-1}$ and $Y^{-1}Z$.

domness may effect the hybridization behavior of polynucleotides. A *general* question is the characterization of those ensembles of strands where the minimum energy diagram(s) can be explicitly described. Another question concerns the hybridization of *diagrams* D_i where the Crick-Watson pairing applies to the remaining single stranded edges of the diagrams. In other words, we consider the space $\mathcal{D}(\{D_i\})$ of the diagrams which can be obtained by the hybridization of the D_i's. This space naturally embeds into the space $\mathcal{D}(\{S_{i,j}\})$, where $S_{i,j}$ are the strands making D_i. In this notation, the problem consists in minimizing the free energy on the subspace $\mathcal{D}(\{D_i\}) \subset \mathcal{D}(\{S_{i,j}\})$; thus, the minimal energy of the diagram hybridization is typically greater than that for the strands making the diagrams. Examples of specified multi-stage hybridization systematically appear in the chemistry of self-assembling DNA nano-devices and DNA computational schemes.

The moduli space of diagrams. Given a collection $\mathcal{S} = \{S_i\}$ of strands, we define the moduli space $\mathcal{D} = \mathcal{D}(\mathcal{S})$ as a directed graph whose vertices are diagrams D made of S_i's and whose edges $D_1 \to D_2$ correspond to hybridization: the diagram D_2 is obtained from D_1 by gluing together a pair of free edges. The moduli space $\mathcal{D}$ has a distinguished source vertex corresponding to the non-hybridized strands, and many sinks corresponding to fully hybridized strands.

The graph $\mathcal{D}(\mathcal{S})$ is almost (modulo possible automorphisms of $\mathcal{S}$) entirely determined by the four numbers $N_A, N_{A^{-1}}, N_G, N_{G^{-1}}$ of the occurrences of the four labels in $\mathcal{S}$, since the Crick-Watson pairing is irrespectful of the topology of $\mathcal{S}$ and of the position of the labels. As a mere graph, $\mathcal{D}(\mathcal{S}) = \mathcal{D}(N_A, N_{A^{-1}}, N_G, N_{G^{-1}})$ looks rather transparent but it comes along with additional (not so transparent) structures. In fact, $\mathcal{D}$ serves as the base space of a "fibration" $f : \tilde{\mathcal{D}} \to \mathcal{D}$ where the fiber $f^{-1}(D) \in \tilde{\mathcal{D}}$ is just the graph $\underline{D}$ underlying D, and where the f^{-1} pullback of each arrow $D_1 \to D_2$ is the cylinder of the corresponding hybridization map between the graphs. Observe that the surface C associated to a diagram D (or rather the actual cylinder of the hybridization map) embeds into $\tilde{\mathcal{D}}$ as a pullback of an upstream path from $D \in \mathcal{D}$ to the source vertex. In particular, the pullbacks of different upstream paths from D to the source vertex are mutually homeomorphic.

The fibration $\tilde{\mathcal{D}} \to \mathcal{D}$ linearizes to the (huge) commutative diagram of the homomorphisms of the homology groups $H_1(\underline{D} = f^{-1}(D))$, where $D \in \mathcal{D}$, with the obvious arrows corresponding to the edges $D_1 \to D_2$. Moreover, each group $H_1(\underline{D})$ comes along with the bilinear pairing (intersection of 1-cycles in the surfaces) in the surface $C_D \supset \underline{D}$. The dimensions of these groups as well as the (co-)ranks of the arrows and pairings are integer value functions on the space $\mathcal{D}$ shaping its combinatorial personality.

Warning: if the collection $\mathcal{S}$ has a non-trivial symmetry (for example it contains two strands of equal length with identical labels, or it contains a circular strand where the labels have a non-trivial period, see Fig. 23) and this symmetry persists under hybridization, one should distinguish between a "diagram" and an "isomorphism class of diagrams". In terms of $\mathcal{D}$, diagrams with symmetries represent singular points where one should specify the corresponding orbispace structure.

Global symmetries do not appear in biologically realistic situations but partial symmetries do appear. In order to incorporate these in the categorical formalism, we need a notion of a *subdiagram* and this can be achieved with the *universal moduli space of diagrams*, $\mathcal{D}^* = \mathcal{D}^*('\infty') = \mathcal{D}^*(\{A, G\})$, which represents diagrams $D(\mathcal{S})$ with variable sets $\mathcal{S}$ of strands, and *injections* $D_1(\mathcal{S}_1) \dashrightarrow D_2(\mathcal{S}_2)$ induced by embeddings $\mathcal{S}_1 \hookrightarrow \mathcal{S}_2$.

Which way to go? Here (and this will be happening over and over again), one faces a dilemma. One can pursue the intrinsic logic of the mathematical construct and follow several mathematical avenues suggested by the space

$\mathcal{D}$: one may investigate its relation(s) to the moduli spaces of Riemann surfaces, one can make a small category out of it and study the topology of its classifying space, one can "complete" $\mathcal{D}$ by allowing diagrams on infinite strands, one may generalize the notion of diagram by replacing the figure $'\infty'$ by a more general (Riemaniann) space (e.g. of negative curvature) or by a suitable homomorphism of the fundamental group of the diagram into another group, etc. Alternatively, one may concentrate on those aspects of hybridization encoded in $\mathcal{D}$, which appear more natural from a bio-chemical standpoint, by trying to find biologically significant patterns in $\mathcal{D}$, by developing computational means to analyze the structure of individual diagrams as well as by introducing new structures on D's and $\mathcal{D}$ suggested by biophysical considerations. The latter leads to new mathematical structures and creates further forks along the road.

Temperature, entropy and Gibbs measures. We have already assigned energy to each vertex $D \in \mathcal{D}$, and now we discuss how to assign a probability (*entropy*) weight to a diagram D. This comes more or less naturally if we "decorate" each D with the space realization of D, that is an embedding $\Phi : \underline{D} \to \mathbb{R}^3$. Such a (D, Φ) is called a *euclidean diagram.* For our purposes, we consider only piecewise linear maps Φ sending edges of D to unit euclidean segments. The "euclidian decoration" of the moduli space $\mathcal{D}$ has three essential ingredients:

topology: the map Φ maybe "knotted", namely there are many isotopy classes of embeddings $\underline{D} \to \mathbb{R}^3$ and this can be recorded by augmenting the combinatorial structure of $\mathcal{D}$.

energy: the combinatorial energy defined earlier can be made more precise by taking into account the actual data on the intermolecular interactions including a realistic bending energy as well as the *hard-core repelling potential* [39] that would automatically force Φ to be an embedding. Observe that the bending energy is local on $\underline{D}$, while the hard-core potential is local on $\underline{D} \times \underline{D}$ rather than on $\underline{D}$, which makes its analysis notoriously complicated (as it reflects the excluded volume or self-avoiding property of Φ).

entropy: this refers to the total measure of the space $E = E(D)$ of euclidean diagrams (D, Φ) as a function of D. The space E naturally embeds into $\mathbb{R}^{3N}$, where N is the number of edges plus the number of vertices of $\underline{D}$.

[39] For two points x, y with $d(x, y) \leq \epsilon$, the "hard-core energy" equals, by definition, $+\infty$.

This space E is invariant under the isometry group of $\mathbb{R}^{3N}$, which makes it infinite and so its measure should be normalized in a suitable way. If we deal with self-hybridization diagrams (obtained by hybridization of a single strand) we can pass to the factor $E/\mathbb{R}^3$. But if disconnectedness of strands and/or diagrams matters, then one should limit the range of Φ to a bounded domain (corresponding to the tube where the solution of DNA is contained). The next problem is that $E \subset \mathbb{R}^{3N}$ is a (semi-algebraic) subvariety of *positive* codimension and one should take special care on how to define the measure on such a subset. One can use the natural induced piecewise Riemaniann measure but it is customary to use the ambient Euclidean measure weighted with the Gibbs factor coming from the energy. In any case, this measure is very difficult to evaluate and one resorts to some approximation to this measure expressed entirely in terms of the combinatorics of D.

Given an energy function $U : \mathcal{D}(\mathcal{S}) \to \mathbb{R}$ and measure weights $\mu(D) \in \mathbb{R}_+$, where $D \in \mathcal{D} = \mathcal{D}(\mathcal{S})$, one is concerned with the pushforward $\nu = U_*(\mu)$ of the measure μ under U. The measure ν on $\mathbb{R}$ is customary represented via its Laplace transform $G(\beta)$, that is the integral of the function $e^{-\beta U}$ over $\mathbb{R} \ni U$ with respect to ν. This equals the *canonical sum*

$$G(\beta) = \sum_{D \in \mathcal{D}} e^{-\beta U(D)} \mu(D)$$

where each term $\mu_G(D) = e^{-\beta D}\mu(D) = e^{-U/T}\mu(D)$ is called the *canonical measure* and where $T = 1/\beta$ is interpreted as the inverse absolute temperature of the ensemble $\mathcal{S}$ of strands in the tube. The canonical sum carries the same information as the specific heat capacity of the ensemble of our strands in solution, which is an experimentally measurable quantity defined as the T-derivative of $\sum_{d \in \mathcal{D}} U(D)\mu_G(D)$. Clearly this derivative equals $(T^2 G'(T))'$.

This canonical sum, or rather its simplified versions, were extensively studied for $\mathcal{S}$ consisting of pairs of complementary strands. In particular, by looking at an Ising type approximation to G, one can make rather realistic predictions of the melting behavior of DNA strands such as the zipping effect. Also, by looking at G restricted to various segments of DNA one finds an amazing matching with biologically significant patterns such as introns and exons in genes, as well as coding regions for functional domains in proteins.

Another biologically significant case, concerns folding of individual strands S of RNA, where "folding" refers to the measure/energy distribution on the diagrams $\mathcal{D}(S)$. One is interested by how much this distribution for S's found in cells differs from that for random S's.

The road to equilibrium. The Gibbs measure does not tell one how the actual hybridization process develops but only describes the final *equilibrium* stage of hybridization. Unlike the statistical ensembles usually studied in physics, the *relaxation time* (i.e. the time needed to reach equilibrium) in biology is relatively long, and the road to equilibrium maybe very bumpy. The non-equilibrium dynamics of ensembles of DNA's can be modeled as a random walk on $\mathcal{D}$, by assigning suitable transition probabilities to the edges between D's in $\mathcal{D}$. The assignment p_{12} to the edge $D_1 \to D_2$ depends on the difference between the Gibbs energies of D_1 and D_2, and reflects to a large extent the combinatorics of the folding (of the map $D_1 \to D_2$ represented by the edge), while p_{21} (which is not supposed to be 0 but usually smaller than p_{12}) reflects the combinatorics of the unfolding.

The major issue is the evaluation of the rate of diffusion of such a walk, or/and of the first eigenvalue of the corresponding diffusion operator. To get the flavor, we localize to the set of diagrams $\mathcal{D}_L \subset \mathcal{D} = \mathcal{D}(\mathcal{S})$, where the total number of double stranded edges equals L. Consider the graph with vertices $D \in \mathcal{D}_L$, where the (non-oriented) edges between $D_1, D_2 \in \mathcal{D}_L$ correspond to the diagrams of oriented edges in $\mathcal{D}$ of the form $D_1 \leftarrow D' \to D_2$, where $D' \in \mathcal{D}_{L-1}$. It seems not hard to evaluate the first eigenvalue of this graph in terms of $\mathcal{S}$ and L, where the total length of the strands in $\mathcal{S}$ and L go to ∞. For example, if $\mathcal{S}$ consists of a single strand S of length $2L$ and $N_A = N_{A^{-1}} = L$ then $\mathcal{D}_L$ can be identified with the Cayley graph of the permutation group Sym_L generated by transpositions. Actually, Sym_L simply and transitively acts on the set of surfaces C_D, where $D \in \mathcal{D}_L$, and appears as a discrete approximation to the Riemann moduli space of surfaces of variable genera. (This example does not incorporate the energy assignment depending on the labelling and cannot tell us much about the random walk we are truly interested in.)

Cleavage, ligation and emergence of stochastic symmetry. Let us bring more chemistry into $\mathcal{D}^*$ by introducing extra edges corresponding to creation and cleavage of covalent bonds between strands. We write $D_1 \Rightarrow D_2$ if D_2 is obtained from D_1 by ligating two ends of two (distinct or identical) strands in D_1. Accordingly, $D_1 \Leftarrow D_2$ signifies breaking a strand at a single location. We assign a weight to each arrow encoding the probability of the chemical

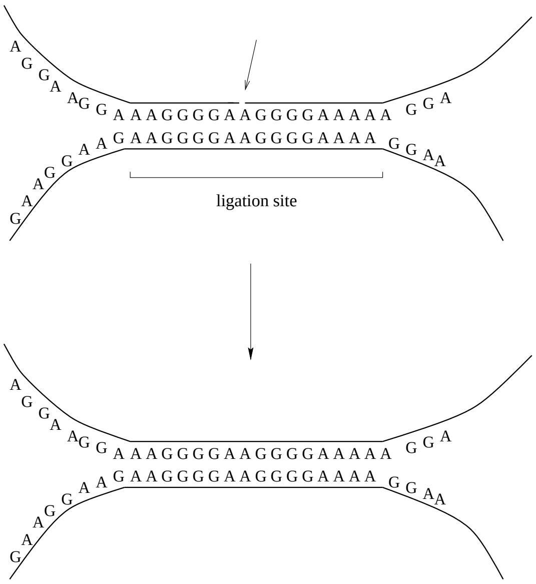

Figure 25: Two strands hybridize with a third one and ligation takes place.

event represented by this arrow. To be realistic, we prescribe a relatively high probability to ligation between two strands hybridized to a complementary strand and having adjacent ends, as in Fig. 16. On the other hand, the probability of ligation of non-adjacent ends is exceedingly low. Somewhat unrealistically, we allow newly broken bonds to ligate again later on. (This presupposes an influx of free energy activating the corresponding nucleotides. Compare Fig. 1.)

We are interested in (random) populations of strands in a tube, where the temperature oscillates between "high", where the strands are completely separated, and "low", where strands conglomerate into diagrams. In the latter case, strands with adjacent ends ligate with relatively high probability. Cleavage happens at all temperatures but with smaller probability.

Question. Consider a random population of strands of various length, i.e. that is a measure on the sequence space. How does this measure evolve in time in the setting described above? Will strands with relatively high repetition of patterns emerge?

Toy DNA symmetrization. To simplify, we work with the two letters A, G

and replace the Crick-Watson complementarity with the pairing AA and GG; we emphasize ligation of two segments S_1, S_2 which are fully hybridized to a third one S_3 such that their ends meet. We think about the subsegment S_3' of S_3 which is hybridized to S_1, S_2 as a "ligation site" for S_1, S_2. See Fig. 25.

A population where subsegments of strands appear with high multiplicity have an advantage as, after cleavage, they have high chance to find hybridization sites. From this angle, the most advantageous (stable) population consists of pure A and pure G strands with a small percentage of interbreeds, since spontaneous A-G ligation appears with probability which is much smaller than spontaneous cleavage, while the A-A and G-G ligations are competitive with cleavage due to the presence of many pure breed ligation sites.

Amazingly, being *random* is sometimes advantageous for reconstruction of cleaved strands, due to the high degree of specification: take a long strand S and a strand S_0' which is twin to a subsegment S_0 of S. In the random case, the binding energy between S_0' and S_0 is roughly twice as great as the energies of other possible bindings of S_0' to S, since the number of matches of letters for random pairs of sequences is roughly equal to the number of mismatches. But for pure breed strands, the binding energy is constant for full hybridization and, in general, proportional to the length of the hybridization overlap. The same consideration applies to the hybridization of two pieces of a cleaved strand on available ligation sites. Since the energy enters the canonical sum under the exponent of the Gibbs-Boltzmann factor, it can beat entropy and, depending on parameters (as temperature, energy, concentration, etc.), the population may evolve towards strands with repeated random motives. One wonders if the tandem repetitions in genomes can be explained by mechanisms of this nature.

What else is there? Mismatches of hybridization make some potential hybridization sites unavailable. In particular self-hybridization of strands (which is predominant at low concentrations) might "tie up" the ends of the strands.

Since the sequence space, where the evolution of strands takes place, is huge compared to the number of strands in the solution, one never arrives at the true equilibrium state in real time. This means that the (limit) equilibrium distribution does not reflect the structure of an actual population at a given moment of time. Instead of staying at equilibrium, one should rather think of a relatively small cloud of strands (*quasi-species* in terms of Manfred

Eigen) wondering in the immense vastness of the full sequence space.

One may try to bring the above model closer to "real life" by taking complementarity and double strands seriously and by introducing other covalent modifications of diagrams, such as recombination for instance. Besides, one can prefer a large pool of random short activated nucleotides, rather than allowing spontaneous reactivation of cleaved strands.

Actual polynucleotides in solution make a complicated ternary structure not easily predictable by the self-hybridization pattern. Some of them display pronounced enzymatic properties, e.g. enhancing cleavage and/or ligation at specific sites of other strands. A possible way to incorporate this phenomenon into the language of diagrams is, following Kauffman, to regard the enzymatic activity as a random function on $\mathcal{D}^*$. Granted such a function, it will significantly influence the evolution process favoring segments which, after folding, promote ligation of self-like, and cleavage of competitors. For example, pure breeds may be at a disadvantage similarly to pure strategies in game theory. It would be interesting to realize this idea in a specific model.

Controlled paths in $\mathcal{D}^*$. There are several experimental techniques to manipulate, control and analyze DNA in solution. These can be described as operations on the space $\mathcal{P}$ of populations of diagrams, that are measures on $\mathcal{D}^*$. Whenever we need *random* populations, we shall refer to the space $\mathcal{R}$. The operations are:

macroscopic compartmentalisation: instead of using one tube, one might have solutions of DNA in n tubes, for relatively small n. This means that we deal with n-dimensional vectors $P = (p_1, \ldots, p_n) \in \mathcal{P}$ of diagram (in particular, strand) populations, where each p_i is a (random) measure on $\mathcal{D}^*$.

mixing solutions: the redistribution of the content of each of the n tubes in a given proportion into m tubes, amounts to a specified stochastic $n \times m$ matrix defining a linear map $\mathcal{P}^n \to \mathcal{P}^m$. If the non-zero weights assigned by p_i to each diagram are sufficiently large (i.e. each diagram appears in sufficiently many copies in solution), this formalism faithfully represents what happens in the lab. But if some diagram appears in a small quantity, the number of them going to a new tube, after mixing can be understood only probabilistically. For example, if we have only one strand of a given species in a tube, we cannot a priori tell in which of the tubes it goes after the mixing process. However, a single molecule matters and can be detected a posteriori, for instance, with

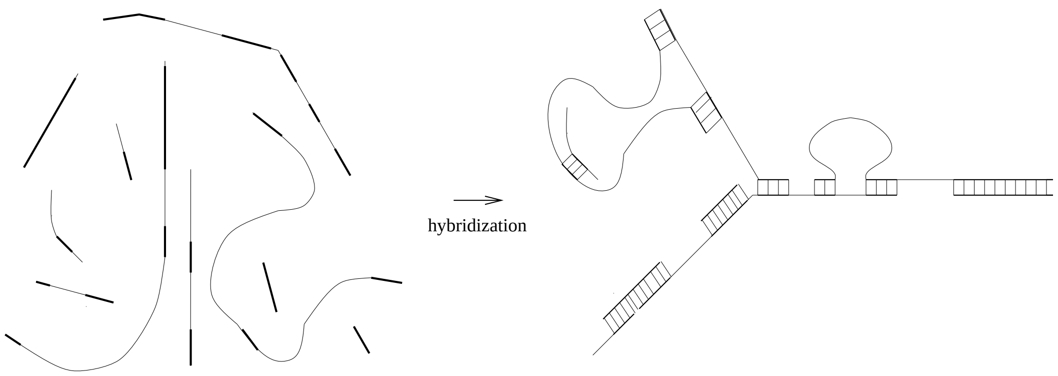

Figure 26: Designed assembly of DNA strands. The bold traits (on the left) correspond to the subwords designed to hybridize (on the right).

amplification by PCR. Thus, we cannot round off small quantities and have to keep track of them using the framework of *random* populations.

temperature control, hybridization and denaturation: raising temperature to sufficiently high degree denaturates DNA and gives us an obvious map from $\mathcal{P} = \mathcal{P}(\mathcal{D}^*)$ to the space of population of strands denoted by $\mathcal{P}(\mathcal{S})$. Conversely, lowering the temperature, leads to hybridization, that is a map $\mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{D}^*)$. If the free energy of hybridization has a unique deep minimum (with a large attraction basin), then the hybridization is essentially unique and non-ambiguous. In general, it is not so and the true map lands in the space of *random* populations of diagrams. Also, the time profile of the temperature may have significant effect on the result of hybridization. For example, if one mixes certain partly hybridized diagrams, they may hybridize along sticky ends arriving at a metastable state corresponding to a deep *local* minimum of the free energy. After heating and slow cooling, one arrives at the *global* minimum providing a population of maximally hybridized strands (which may appear in different knotting/linking conformations).

Hybridization can be designed in order to serve (at least) two purposes: the *controlled self-assembly* of a diagram out of strands and the realization of *non-deterministic algorithms*. In the first case, one needs the free energy with a unique sharp pronounced minimum, and in the second case one seeks free energy functions with several deep (local) minima with equal values of the free energy. These two purposes are achieved by taking a collection of randomly labelled strands modulo the following condition: there is a distinguished relatively small set of mutually complementary words which appear as subwords in our strands. (See Fig. 26.) Then, the minima of free energy are realized by diagrams hybridizing across these words. In both cases, one starts with many copies of each strand. For self-assembly purposes, one needs to purify the resulting population of diagrams in order to exclude undesirable hybridizations. To realize a non-deterministic tree with N leaves, one needs as many copies of each strand as is necessary to realize N different minima with non-zero probability.

making primers: this means construction of specified vectors of measures in $\mathcal{P}$ or in $\mathcal{R}$, supported on the space of relatively short strands.

separation: this is a map $s = s_\pi : \mathcal{P} \to \mathcal{P}^N$, where π is a partition of $\mathcal{P}$ into N subsets corresponding to the physical characteristics involved. The map s_π consists of restricting a measure to each, among N, subsets of the partition. The partitions π that we have in mind are those which can be biochemically realized. For example, one can separate strands with gel-electrophoresis according to the length, or diagrams according to their mass and overall topology.

purification: it is the multiplication of a measure by the characteristic function of some subset in $\mathcal{D}^*$, interpreted as filtering away the diagrams where the function vanishes. This should be *physically* realizable.

extraction: this is a map $ex = ex_{\pi,r} : \mathcal{P} \to \mathcal{P}^N$, where π is a partition of $\mathcal{P}$ into N subsets, $r = \{r_i\}_{i=1}^N$ are positive weights ≤ 1, and where the i-th component of $ex(p)$ equals p restricted to the i-subset of the partition times r_i. Formally speaking, extraction can be reduced to separation, purification and mixing, but, in practice, a particular extraction may be feasible while the corresponding separation and/or purification are not necessarily so. For example, one can extract those diagrams which have a non-hybridized segment(s) with a given labelling(s) using complementary probes attached to a solid support.

detection of $p \in \mathcal{P}$: this refers to the evaluation of the integral $p(d)$ of a given function $d : \mathcal{D}^* \to \mathbb{R}$. We use those d's where this evaluation is experimentally feasible, preferably even for small values of $p(d)$. An example of this is provided by molecular beacons, where the function d take values $0, 1$ depending on the presence or absence of a given subword in a strand, and the intensity of luminescence equals the integral (average) of this (characteristic) function.

changing connectivity of $\mathcal{S}$: an enzyme is a map $e_i : \mathcal{D}^* \to \mathcal{D}^*$ which breaks or ligate strands at given locations, which are specific for a given enzyme. We admit enzymes making several cuts within the same location (for example EcoRI makes two cuts, as illustrated in Fig. 19). Thus every enzyme is characterized by the size of the site at which it can work, by the combinatorics and labelling of the site, and by the combinatorics of the site after the action of the enzyme. Given e_i, we obtain a map $E_i : \mathcal{P} \to \mathcal{P}$ by linearly extending e_i.

disactivation of ligation: one can disactivate the ends of the strands in a given tube (compare to the paragraph on *Synthesis of polynucleotides* in Section 5) and thus prevent the strands from ligation. The disactivation is a labelling of certain ends of strands that allows a design of a wider variety of ligation maps e_i.

complementation: given a diagram, one can generate from it the complements to all non-hybridized segments using some replicase enzyme in the presence of an excess of free nucleotides. This amounts to adding the formal sum of these subsegments to the δ-measure supported on this diagram.

Also, one may generate a specific subsegment of each non-hybridized segment, starting from a given short subword in it and using a primer complementary to this subword together with DNA polymerase enzymes. This amounts to an operation on populations similar to the above.

amplification: repeatedly applying complementation and denaturation leads to measures on $\mathcal{D}^*$ having an exponentially large weight on specific strands. In particular, with replicase, one can multiply a given measure, symmetric under complementation, by 2^n in n steps.

Granted these operations with specified parameters for the free energy and error margins expressed in the degree of randomness of the arising diagrams, one poses two questions:

- *can one faithfully assemble a particular diagram?* Here, one is additionally interested in a diagram with a prescribed embedding into $\mathbb{R}^3$, e.g. in the realization of specified DNA knots. (Many classes of single and double stranded knots have been already chemically implemented.) For this, one has to modify the above formalism by incorporating in it the combinatorics of the double helical structure of DNA;

- *which (non-deterministic) computations can be modeled by the above operations on population of diagrams?* For example, given two populations p_1, p_2, how many operations one should perform over p_1 in order to obtain p_1' approximating p_2 with a prescribed error?

Lesson from $\mathcal{D}^*$. Our formal description of DNA is by no means final neither from a bio-chemical nor from a mathematical standpoint. It is not sufficiently specific for practical applications; yet, it illustrates some of the general principles mentioned earlier: complementarity and templating, self-assembling where hybridization and denaturation serve as scaffolding for making/breaking covalent bonds, repetitiveness of stochastic motives, channeled relaxation via designed free energies.

On the mathematical side, one seeks for a more general class of models incorporating random walks on combinatorial moduli spaces and the design of (free energy) functions with specified symmetry of local minima with an explicit separation of several (at least three) scales. (In the $\mathcal{D}^*$-model the weak bonding serves as intermediate between thermal energy and covalent energy, where the concentration of activated bases appears as "covalent temperature".)

SELF-ASSEMBLY.

Tilings. Consider a connected subset T (*tile*) in $\mathbb{R}^3$, for example a convex polyhedron, with a distinguished subset of mutually complementary (possibly overlapping) non-empty domains on the boundary, denoted $D_b, \bar{D}_b \subset \partial T$, where b runs over a (possibly infinite) set B. We are interested in *assemblies generated* by T, that are subsets A in the Euclidean space, decomposed into a union of congruent copies of T where two copies may intersect only at their boundaries and have a "tendency" to meet across complementary domains on the boundary. We have in mind a protein molecule T with complementary active sites $D_b, \bar{D}_b$ such that different copies of T bind along the complementary domains and self-assemble into complexes. In the geometric context we specify the binding properties by introducing *(binding) isometries* $b : \mathbb{R}^3 \to \mathbb{R}^3$ to each $b \in B$ such that T and $b(T)$ intersect only at the boundary, and $b(D_b) = \bar{D}_b$. From now on B is understood as a subset in the eucledian isometry group $Iso(\mathbb{R}^3)$.

Accordingly, we define an *assembly A associated to* (T, B) by the following data:

- a connected graph $G = G_A$ with the vertex set $1 \dots N$,

- subsets T_i in $\mathbb{R}^3$, where $i = 1 \dots N$, which may mutually intersect only at their boundaries,

- an isometry $b_{k,l} : \mathbb{R}^3 \to \mathbb{R}^3$ moving T_k onto T_l, for each edge (k, l) in G, such that there exists an isometry $a_{k,l}$ which moves T_k to T and conjugates $b_{k,l}$ to some isometry b in B. Notice that this b is uniquely determined by $b_{k,l}$ up to conjugation.

Several tiles. If we start with several different tiles $T^1, \dots, T^n$ rather than with a single T, we consider the sets of pairs of binding isometries $B^{i,j} \subset Iso(\mathbb{R}^3) \times Iso(\mathbb{R}^3)$ such that $b_1^{i,j}(T^i)$ and $b_2^{i,j}(T^j)$ intersect only at their boundaries and their intersection is non-empty. The definition of an assembly associated to $(\{T^i\}, \{B^{i,j}\})$ goes as above with the following modifications: the graph G has vertices colored by the index set $1 \dots n$, the corresponding subsets in $\mathbb{R}^3$ are denoted T_k^i where $i = 1 \dots n$ and $k = 1 \dots N_i$, and finally, we forfeit the isometries $b_{k,l}$ and for each edge (k^i, l^j) we emphasize an isometry of $\mathbb{R}^3$ which moves T_k^i to $b_1^{i,j}(T^i)$ and T_l^j to $b_2^{i,j}(T^j)$.

In what follows, we sometimes refer to the union of tiles defined above, as an *assembly.*

Qualities of an assembly. The *tightness* of the tiling is one quality that chemists appreciate. This can be measured by the number of cycles in the graph G, or equivalently by the minus Euler characteristic of the graph.

The *imperfection* of a tiling is measured by the "unused" areas of the boundaries of the tiles. First define the *active domain* $\partial_{act}(T) \subset \partial T$ as the union of the intersections of ∂T with $b(T)$ for all $b \in B$. Then define the "unused boundary" $\partial_{un}(A = \cup T_i)$ as the union $\cup_{i=1}^N \partial_{act}(T_i)$ minus the union of the pairwise intersections $\cup_{(k,l) \in G} T_k \cap T_l$. An assembly is called *perfect* if the area of the imperfection equals zero. We say that an assembly contained in a *given* subset $X \subset \mathbb{R}^3$ is *perfect* with respect to ∂X, if $\partial_{un}(A) \subset \partial X$.

The *uniqueness* refers to the uniqueness of an assembly subject to some additional constraints. For example, given an $X \subset \mathbb{R}^3$, one asks first if X can be tiled by (T, B) and then asks for the uniqueness of such a tiling. We say that (T, B) generates an *unconditionally unique* assembly if every imperfect assembly uniquely extends to a perfect assembly.

Figure 27: Vernier. Rodlike tiles differing in length form an assembly that grows until the ends exactly match.

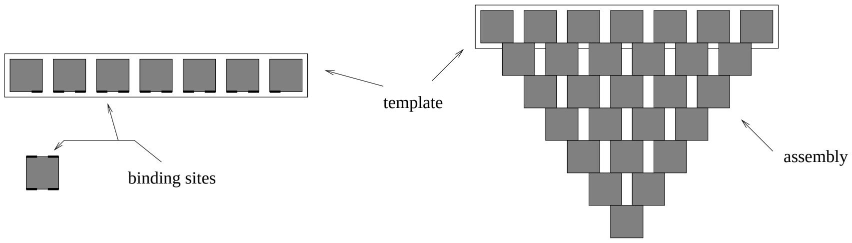

Figure 28: A tile is stable in the assembly only if it binds at two adjacent binding sites. The stability of the whole assembly is insured by the enforced stability of the template. The formal description of this example is not completely captured by our model.

The essential problem of tiling engineering is designing a relatively simple tile or a few such tiles which assemble with high quality into large and complicated subsets in $\mathbb{R}^3$. Here is a specific example for the unit sphere S^2 rather than S^3, where one uses the obvious extension of the notion of tilings to homogeneous spaces. Given $\epsilon, \delta > 0$, consider triangulations of the sphere into triangles Δ with $Diam(\Delta) \leq \epsilon$ and $area(\Delta) \geq \delta Diam^2(\Delta)$. It is easy to see that the number of mutually non-congruent triangles in such a triangulation, call it $n(\epsilon, \delta)$, goes to ∞ for $\epsilon \to 0$ and every fixed $\delta > 0$. The problem is to evaluate the asymptotic behavior of $n(\epsilon, \delta)$ for $\epsilon \to 0$ and either a fixed δ or $\delta \to 0$.

Real life examples. It remains unclear, in general, how cells control the size of (imperfect, with some unused boundary,) assemblies, but certain mechanisms are understood. For example, out of two rodlike molecules of length three and five, one gets a double rod of length 15 as illustrated in Fig. 27. Another strategy is starting an assembly from a given template (see Fig. 28 for a specific design). Sometimes, tiling is non-isometric: tiles slightly bend in order to fit, and the assembly terminates when the bending energy becomes too costly or when the accumulated bending deforms and disactivates the binding sites (see Fig. 29). Also, the binding of a ligand to an active

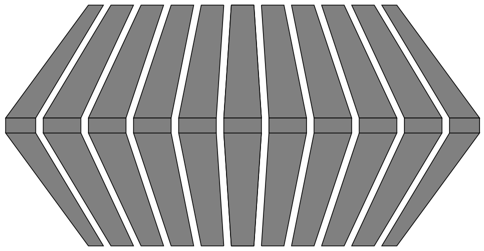

Figure 29: Polymeric structure growing until the energy required to fit new subunits becomes too large.

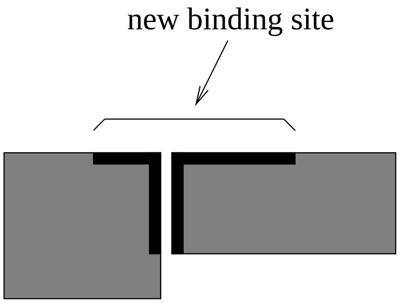

Figure 30: Tiles which differ in shape and binding sites. Their binding generates a new contiguous binding site.

site might change the shape of the molecule and thus influence the binding activity of other sites. Another possibility is the creation of a new binding site distributed over two or more tiles bound together on an earlier stage of the assembly (see Fig. 30).

These mechanisms may produce a non-trivial dynamics in the space of assemblies in the presence of free-energy. (The delivery of free-energy into a self-assembling system is hard to realize bio-chemically. Even mathematically, modeling such processes does not seem easy.) In particular, one may try to design a periodic motion of a tile over a template, something in the spirit of a RNA-polymerase cycling around a plasmid.

Programmed design. Can one find a small set of relatively simple tiles such that, starting from a template supporting a linear code (that may be a DNA or RNA molecule incorporated into a macromolecular complex), the assembly process will create a given three dimensional shape in the space? We think here of interacting tiles performing a transformation from labelled templates into three dimensional structures and we ask what kind of transformations can be realized in this way. Also, we want to understand how much the complexity of the construction depends on the complexity of the tiles, where the latter can be measured by the number of the binding sites of the tiles, the size of the sets $B^{i,j}$, etc.

Liposomes and minimal surfaces. Consider a smooth closed surface $S \subset \mathbb{R}^3$ and let S_δ denote a small normal δ-neighborhood of S, that is the union of the 2δ-segments $[-\delta,\delta]_s \subset \mathbb{R}^3$, over all $s \in S$. We want to think of S_δ being tiled by the segments $[-\delta,\delta]_s$ and to make this realistic we replace the segments by thin rods as follows: pack the surface S by N ϵ-discs $D_{s_i,\epsilon} \subset S$, where $i = 1 \dots N$, $s_i \in S$ and where ϵ is much smaller than δ; let $l_{s_i,\epsilon,\delta}$ be the union of the segments $[-\delta,\delta]_{s'}$ over $s' \in D_{s_i,\epsilon}$. The union $L = L_{S,\epsilon,\delta} = \cup_{i=1}^{N} l_{s_i,\epsilon,\delta}$ is called the *micelle* assembled out of the rods $l_{s_i,\epsilon,\delta}$.

Such "micelles" appear in nature, in particular in cellular membranes, where the major structural component is a phospholipid bi-layer built of rod-shaped molecules having one hydrophobic and one hydrophilic end. Such bi-layer surfaces, called *liposomes*, can assemble spontaneously in a solution containing phospholipids. See Fig. 31. The shape of a liposome is determined by the equations governing hydrophobic forces which transform, in the limit for $\epsilon, \delta \to 0$, to specific partial differential equations on the resulting surface S. For example, the membrane of an erythrocyte (i.e. red blood cell) is

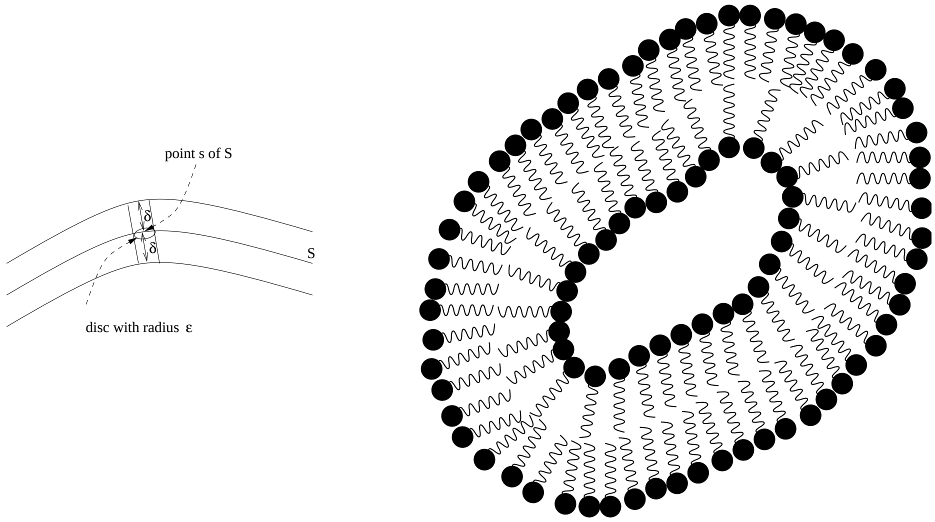

Figure 31: The schema of a rod lying along a surface S (left), and the section of a liposome, a bi-layer surface packed with rod-shaped molecules (right).

believed to minimize its integrated squared curvature with a given area, and enclosing a given volume [40]. This suggests the following:

Question. Given an isotopy class of closed surfaces in $\mathbb{R}^3$ does there exist a surface S in this class with a given area and a given enclosed volume which minimizes the integral of a given function of principal curvatures? (The most studied case concerns the integrated squared mean curvature as it enters the Willmore conjecture; other integrants were not apparently investigated much.)

Tilings, energies and variational equations. Let us encompass the above into a more general scheme incorporating energy into the idea of assemblies of tiles.

Let V be a smooth manifold ($V = \mathbb{R}^3$ in the previous discussion) and let $\mathcal{P} \to V$ be a smooth fibration associated to the full frame bundle $Fr \to V$. We think of the fiber $P_v \subset \mathcal{P}$, for $v \in V$, as the configuration space of a tile (a protein molecule) in V located at v. In fact, the position of a tile

[40]This is assumed by the bio-medical community but apparently there is no mathematically rigorous proof. The main technical point is to establish the axial symmetry of the solution of the variational problem.

in the Euclidean space is uniquely determined by the location $v \in \mathbb{R}^3$ of its center of gravity and an orthonormal frame at this point, that is a 3-tuple of orthonormal tangent vectors at v. Similarly, a phospholipid molecule is modelled by a tangent vector in $\mathbb{R}^3$.

The configuration space for N identical tiles is the Cartesian power $\mathcal{P}^N$, where the rules of tiling are encoded into a (energy) function $U : \mathcal{P}^N \to \mathbb{R}$. A distinguished class of energies is given by the two-particle interactions, that are functions $u : \mathcal{P}^2 \to \mathbb{R}$, with U defined in the usual way:

$$U(p_1, \ldots p_N) = \sum_{i,j=1}^{N} u(p_i, p_j).$$

Example. In the language of the paragraph on *Tilings* (see earlier in this section), the function u is defined on the pairs of frames (p_1, p_2) in $\mathbb{R}^3$ representing pairs of tiles (T_1, T_2) as follows:

- if the interiors of the tiles T_1 and T_2 intersect, then $u(p_1, p_2) = +\infty$.

- if T_1 intersects T_2 across binding domains on their boundaries defined with $b \in B$, then $u(p_1, p_2) = u_0(b)$ for a given negative function $u_0 : B \to \mathbb{R}$ recording the corresponding binding energy.

- if the tiles T_1 and T_2 are disjoint then $u(p_1, p_2) = 0$.

Observe that this potential is singular and discontinuous. A natural class of potentials is constituted by those which are smooth on strata of some stratification of the configuration space. Often, one needs to regularize them and keep track of the fine local geometry of the stratification.

One can define tilings/assemblies as local minima of U. Everything we said about tilings before can be reformulated in this language. Physically speaking, this (local) minima correspond to (meta)stable states at zero temperature. In order to bring in positive temperature, we need a canonical measure on the space $\mathcal{P}$ and thus, on the space $\mathcal{P}^N$. Usually, for instance for frames and rods in the Euclidean space (and in Riemaniann manifolds V in general), there is a non ambiguous canonical measure invariant under the Euclidean isometry groups. In most cases, the choice of the canonical measure is influenced by the Louiville measure on the phase space "overlying" $\mathcal{P}^N$.

Given a canonical measure $d\mu$, one can speak of Gibbs states (measures) $e^{-\beta U} d\mu$ and model an approach to equilibrium (Gibbs) states by biased random walks in $\mathcal{P}^N$. The quality of a tiling can be now expressed as the rate of approach to equilibrium and the size(s) of the attraction basins of various states. Much of the discussion on population of diagrams extends to population of tilings.

Minimal surfaces. Given a class of subvarieties $W \subset V$ distinguished by a variational equation, one wants to obtain them as limits of Gibbs states with suitable U on $\mathcal{P}^N$, for $N \to \infty$, with simultaneous rescaling of U. Is it always possible? Can one go away with two-particle potentials? (The idea is to divide a minimal surface into "particles" corresponding to infinitesimal elements, dissolve them in a tube and then reassemble them, following a biased random walk guided by some energy function between the particles. We want the reassembled object to satisfy the original partial differential equations.) The latter seems unlikely in view of the fact that the hydrophobic forces shaping liposomes are not two-particle [41]. (If one thinks of the limits of N-tuples $\{p_i\}$ as measures on $\mathcal{P}$, one is lead to a generalization of Almgren's theory of varifolds where the potential is not reducible, or at least not immediately, to interactions between finitely many particles.)

Bending, curvature and multi-jets spaces. In order to incorporate bending and the resulting energy of molecules, one can generalize the above discussion by replacing Fr by the k-jet frame bundle Fr_k which, for example, encodes curvatures for $k = 2$. This suggests looking at higher order variational equations, where basic examples are provided by complex algebraic geometry.

Given a complex submanifold W of dimension m in a Kähler manifold V, one can associate to it submanifolds $W_{k,N}$ in larger Kähler manifolds by taking N-tuples of suitably reduced k-jets of germs of W in V. Each of these $W_{k,N}$ is volume minimizing in the ambient (complex) space (which comes with a natural Kähler metric), and thus the original W satisfies a hierarchy of multi-differential equations. Can these equations be recaptured from appropriate Gibbs states? How much of the complex analytic nature of W is reflected in the (diffused) Gibbs states? Probably one needs here a multiscale

[41] Probably, one can formally reduce hydrophobic forces to two-particle interactions by explicitly introducing the water molecules into the picture. Here, one needs to go to the double limit, $size(H_2O)/\delta \to 0$ and $\delta \to 0$, expressing the idea that the water molecules are small compared to the size δ of phospholipids.

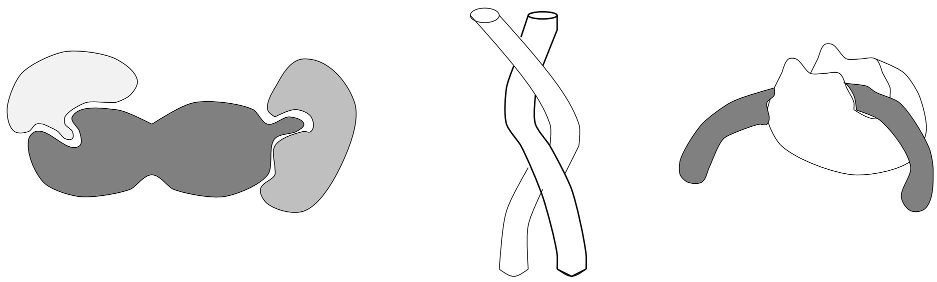

Figure 32: Protein complexes formed by binding of complementary rigid surfaces (left), by helix-like coiling (center), and by binding of a long polypeptide chain with a rigid surface (right).

limit where the energy comes on different levels (weak bonds, covalent bonds, etc.) and takes care of different derivatives and/or singularities.

Complementarity, affinity and protein design.

An approximate complementarity relation on a set P (which may be a set of proteins or of macromolecules in general) is a positive symmetric function $a : P \times P \to \mathbb{R}$, where $a(p_1, p_2)$ measures a certain affinity between p_1 and p_2. One thinks of the relation $A_\epsilon = \{(p_1, p_2) | a(p_1, p_2) \geq \epsilon\} \subset P \times P$ as a multivalued map $A_\epsilon : P \to P$ and one wants to express the idea of this map being an approximate involution. This is done with the inequality

$$(*) \qquad a(p_1, p_4) \geq F(a(p_1, p_2), a(p_2, p_3), a(p_3, p_4))$$

where $p_i \in P$, for $i = 1 \dots 4$, and $F = F(a_1, a_2, a_3)$ is a positive symmetric function monotone increasing in each variable (chosen accordingly to a given situation). Such an F is supposed to capture the idea of the approximate *key-lock* binding mechanism between proteins and other (macro)molecules. See Fig. 32 for possible geometric patters of binding mechanisms.

Examples. Let P consist of the polynucleotides of a given length l and let a be the energy of the best hybridization between them, where we assume that the energy of the hybridization equals the number of mutually complementary pairs at the corresponding positions $j = 1 \dots l$. We write a polynucleotide as a string $p(j)$ and agree that $p_1(j) \cdot p_2(j) = 1$ if the bases $p_1(j)$ and $p_2(j)$ are complementary, and 0 otherwise. We express the hybridization energy as

$$a(p_1, p_2) = \langle p_1, p_2 \rangle = \sum_{j=1}^{l} p_1(j) \cdot p_2(j).$$

Clearly, $(*)$ is satisfied here with $F = max\{0, a_1 + a_2 + a_3 - 2l\}$. This is meaningful only for large a's, namely $\geq 2/3$. On the other hand, the affinity between two random strands equals $1/4$ of l. Furthermore, if p_i, for $i = 1 \ldots 4$, are random strings conditioned by the inequalities $a(p_i, p_{i+1}) \geq a_i$, for $i = 1, 2, 3$, then there is a better (easily computable) lower bound on $a(p_1, p_4)$. This suggests that it may be more useful for practical problems (see below) to restrict a to *random* points $p \in P$ where $(*)$ is satisfied with a greater F.

Design of antibodies. Given $p_0 \in P$, one seeks $p \in P$ with possibly large, ideally maximal, value of the affinity function $a(p_0, p)$. In practice, p_0 is a given macromolecule, e.g. a protein, and one looks for a protein p which fast and strongly binds to p_0. "Strongly" usually refers to the *dissociation time* T_{dis} between p and p_0, that is the half life of the association between p and p_0. "Fast" is indicative of the *association time* T_{ass}, i.e. the average time needed for forming the bound state from the unbound one. In practice, one does not work with only two single molecules p_0 and p but rather with ensembles of them in solution, and the times T_{dis}, T_{ass} are measured per given concentrations.

Ideally, one wishes an antibody to be both strong and fast. Pharmacologically, one is happy with p with a relatively short (but not too short) dissociation time and the shortest possible association time. (Early immune response is crucial for fighting an invader.)

Let X be the space representing pairs of molecules p_0 and p in solution (where the space scale is adjusted to implement a realistic concentration of ensembles of molecules in solution: two particles ask for a very small tube!). The ratio of the times T_{dis}/T_{ass} can be expressed in thermodynamical terms as the ratio $C_\propto$ of the Gibbs measure on X of the bound states ($p_o \Bumpeq p$) and the Gibbs measure of the unbound states ($p_0 \not\asymp p$)

$$T_{dis}/T_{ass} = C_\propto = \mu(X_{\Bumpeq})/\mu(X_{\not\asymp}),$$

since the space average equals the time average. This ratio depends on the binding energy of p_0 to p and can be, in principle, computed provided that we know the geometry/energy profiles of the binding sites of the molecules p_0 and p.

The second relevant thermodynamical invariant, the average *transition rate* $R_\propto$, refers to the "area" of the common boundary $X_\propto$ between $X_\Bumpeq$ and $X_{\not\asymp}$ in X. This depends not only on the Gibbs measure but also on the dynamics in the space X describing the time behavior of the molecules and equals the average number of intersections of time orbits with $X_\propto$. An alternative geometric definition is ϵ^{-1} times the Gibbs volume of the ϵ-orbit of $X_\propto$, for $\epsilon \to 0$. (This makes sense for stochastic as well as deterministic time behavior and we shall return to this later.) The times T_{dis} and T_{ass} are expressible in terms of $C_\propto$ and $R_\propto$ as follows:

$$T_{dis} = \frac{C_\propto}{(C_\propto + 1) \cdot R_\propto} \qquad \text{and} \qquad T_{ass} = \frac{1}{(C_\propto + 1) \cdot R_\propto}$$

The theoretic determination of $R_\propto$ for proteins is more difficult than that of $C_\propto$ because it depends on the effects of *weak* affinity over the whole surfaces of the proteins, and not only at their binding sites. On the other hand, the times T_{dis} and T_{ass} can be measured experimentally with essentially equal ease (or difficulty).

The affinity function a extends by bilinearity to the space Π of populations of proteins, that are measures π on P. The experimental techniques provide a bonus of measuring $a(\pi_1, \pi_2)$ essentially with the same ease as $a(p_1, p_2)$: given two populations π_1, π_2 in two tubes, one mixes the contents of these two tubes and isolates the pairs of proteins which stuck together. In practice, the proteins from the population π_1, that are the antigene(s) for which we seek an antibody, are attached to a solid support, and used to extract antibodies p's from π_2 that are free in the solution. The latter are the antibodies which are displayed for example on the coating of viruses (as mentioned at the end of Section 5). This allows one to measure not only the value of $a(\pi_1, \pi_2)$ but also to identify *specific* proteins in the populations π_1, π_2 with mutual high affinity.

The computational power of such an experiment is quite impressive: with 10^5 proteins on the solid support and 10^5 in the solution (these amounts are realistic) one achieves a selection among 10^{10} possible pairs of proteins.

This procedure is repeated many times using artificial evolution of viruses selected according to affinity of the corresponding antigene(s). Mathematically speaking, one maximizes the function $a(p_0, p)$ by the following algorithm: start with an arbitrary p which hopefully gives a relatively large value to $a(p_0, p)$, and then take a population of small ($\approx 10^5$) variations p_i of p. Choose among the p_i's, the one with the largest $a(p_0, p_i)$ and repeat the procedure with this p_i instead of p.

One can make this process more efficient with the same computational effort by varying p_0 as well as p. For example, at the i-th step one can replace p_0 with a population $\pi_0(i)$ of proteins, where $\pi_0(i)$ converges to p_0 for $i \to \infty$, and apply the i-th step to $a(\pi_0(i), p)$. This is a schematization of the approach being currently implemented by Bill Huse in collaboration with Michael Freedman (personal communication by Huse) [42]. One wonders whether the approximate complementarity of the antibody/antigene interaction expressed by $(*)$ can be of any use for protein engineering.

States, dynamics and relaxation.

There are several dialects in the language of dynamical systems. We shall limit ourselves to three of them: *point dynamics*, *fuzzy dynamics* and *stochastic dynamics*. To get the idea, start with a Riemaniann manifold V where:

- *point states* are tangent vectors and the dynamics is given by the geodesic flow;

- *fuzzy states* are subsets $U \subset V$ and the image after the dynamics t-map is the t-neighborhood of U, that is the set of point in V within distance $\leq t$ from U. Alternatively, one may restrict to the boundary ∂U transformed to the boundary of the t-neighborhood of U, following, apart from the wavefront singularities, the geodesic flow applied to Legendrian submanifolds in the unit tangent bundle.

- *stochastic states* are (probability) measures on V and dynamics is the common diffusion semi-group defined by the heat kernel.

Following the classical physics paradigm, one looks for *point* dynamics describing real life systems. If necessary, the space is enlarged by introducing extra "hidden parameters". In the cell, however, the "complete states" carrying the full information for the time development, are apparently not points but probabilities on the configuration space X of atoms and molecules

[42] Another context where the idea of introducing parameters appears is the Shub-Smale algorithm for finding zeros of complex polynomials.

constituting the cell. (It seems unreasonable to keep track of momenta of particles; some quantum parameters of the molecular states though, may be relevant for the functioning of the cell.)

The space X is too large and too fine to be directly observable and what one sees in experiments are states in some reduced spaces which are certain quotient spaces of X [43] defined by classes of phenomenological observables. This cannot be explained without resorting to formal definitions.

A *point predynamics* on X is a map $A : X \times T \to X$, where T is the time domain which is usually either $\mathbb{R}$ or $\mathbb{R}_+$. It might become a more general semi-group, if needed, such as $\mathbb{R}^n$ for Cartesian products of n systems with a possible symmetry reduction due to permutations between identical particles [44].

A *fuzzy predynamics* on X is a map A with X replaced by the space $\mathcal{P}(X)$ of subsets of X, or by some "reasonable" subspace in $\mathcal{P}(X)$, e.g. *measurable* subsets, *closed* subsets, *semi-algebraic* subsets, Lagrangian and Legendrian submanifolds, etc. One usually requires $A(X_1 \cup X_2, t) = A(X_1, t) \cup A(X_2, t)$ and $A(X_1 \cap X_2, t) \subset A(X_1, t) \cap A(X_2, t)$, for $X_1, X_2 \subset X$.

Similarly, *linear stochastic predynamics* is defined with the space $\mathcal{M}(X)$ of measures on X instead of $\mathcal{P}(X)$, where we usually restrict to Borel measures μ on topological spaces X (or to smooth measures on manifolds X) and we require A to be linear in $\mu \in \mathcal{M}(X)$.

A point predynamics on X naturally induces the fuzzy and the stochastic predynamics.

A predynamics A is called *dynamics* if it satisfies the semi-group property: $A(A(x, t_1), t_2) = A(x, t_1 + t_2)$.

Example. Let X be a metric space and A be a fuzzy predynamics, where $A(Y, t)$, for $Y \subset X$, equals the t-neighborhood of Y. If X is a Riemaniann manifold or an arbitrary length space for this matter, then this predynamics is a dynamics. Conversely, if this A is a dynamics then X is a length space (modulo trivial readjustment). Thus, a fuzzy dynamics structure on a space generalizes the (non-symmetric) length metric structure. (One can *meaningfully* generalize further, by allowing a multi-dimensional T. Compare

[43] There are other mechanisms of reduction besides naïve quotients, such as the *symplectic reduction* for example, but these lie out of the scope of the present article.

[44] The classical geometric situation of "interesting" time presents itself in the Weyl chamber flows over real and p-adic locally symmetric spaces.

footnote 44.)

The Gauss' shortest path construction of the intrinsic metric out of the extrinsic one, suggests how to go from a predynamics A to a dynamics $\overset{\rightsquigarrow}{A}$ in general:

$$\overset{\rightsquigarrow}{A}(x,t) =_{def} \lim_{i\to\infty} A^{(i)}(x, i^{-1}t).$$

Here, $A^{(i)}$ stands for the i-th iteration of A, and we apply this definition to $T = \mathbb{R}_+$ whenever the limit exists. (In the physical tradition, every limit exists unless otherwise proven.)

Involutive dynamics. A point dynamics is called *involutive* [45] (time reversible) if there exists an involution on the space X reversing the direction of the dynamics

$$A(Invo(x), t) = Invo(A(x, -t)).$$

For example, the geodesic flow is obviously involutive as well as the system of classical mechanics described by second order differential equations.

A fuzzy dynamics is *involutive* if

$$(A((A(Y,T))^{\perp}, t))^{\perp} \supset Y$$

where $Y \subset X$ and $\perp$ is the operation of complementation of subsets of X.

Riemaniann geometry is involutive as the metric is symmetric. The non-strictness of the displayed inclusion is due to the presence of wavefront singularities.

A stochastic dynamics is *involutive* if the infinitesimal generator Δ of A, provided it exists, is a symmetric operator. This is the case, for example, for the diffusion on Riemaniann manifolds where the symmetry of the (Laplace operator) Δ results from the symmetry of the metric. Similarly, in the chemical kinetics, the involutive property (*detailed balance*) of the underlying point dynamics implies that for the stochastic dynamics (*Onsager relations*).

Reduced dynamics. Given a surjective (factorization) map $f : X \to V$, we want to derive a dynamics on V, representing the space of phenomenological states, from a given A on X. Every fuzzy (in particular, point) dynamics on X obviously projects to the fuzzy predynamics on V. Sometimes it needs to be transformed to a dynamics by the above procedure, but sometimes

[45] Flows of Weyl chambers suggest larger reflection groups compatible with dynamics.

it comes for free. (As it happens, for example, to the geodesic dynamics projecting from the tangent bundle to the underlying Riemaniann manifold.)

The reduction of stochastic dynamics needs distinguished (canonical) measures, one on X and one on V [46], so that certain measures, called *smooth*, are representable by density functions. Such a function can be lifted from V to X, turned into a measure, transformed by A and then pushed forward back to V. This brings down a stochastic dynamics from X to V under a suitable smoothing effect of A on measures.

If we start with an involutive point dynamics on X and the map f is *involutive* (i.e. $f(Invo(x)) = f(x)$), then the resulting fuzzy dynamics is involutive. This is also true in the stochastic case if (in the agreement with the Onsager relations) the point dynamics preserves the canonical measure and this measure pushes forward to V.

In some cases, the full space of point states can be reconstructed from its phenomenological reduction(s) V in a meaningful way. For example, if V is a Riemaniann manifold or a general length metric space then one may take the space X of all locally isometric maps from $\mathbb{R}$ to V with the obvious action of $T = \mathbb{R}$. Clearly, the symmetry of the metric make the geodesic dynamics involutive. If the space V is smooth (or more generally, of curvature bounded from below), then the geodesics cannot branch. This corresponds to causality in a physical situation. On the other hand, the branching of geodesics at the point of infinite negative curvature is reminiscent to the decay of a molecule in the course of a chemical reaction. The latter process seem to need an infinite dimensional space X even if we start with a finite dimensional space V, due to infinite repetitions of the ambiguities involved in the decay process. Here is a construction of X showing this phenomenon. Let A_1, A_2 be point dynamics on spaces V_1, V_2 respectively, and let W be a multivalued correspondence between V_1 and V_2 which is given by a pair of "fibrations" $W \to U_1 \subset V_1$ and $W \to U_2 \subset V_2$. Define X as the space of orbits constructed as follows: take a point $v \in V_1 \cup V_2$, say in V_1, and follow it by the A_1 dynamics until its first entry $u_1 \in U_1$; then take a point $u_2 \in U_2$ *corresponding* to u_1, which means that u_1 and u_2 come from the same point in W. Continue by applying A_2 to U_2, until the orbit returns to U_2 and repeat the same process indefinitely.

This construction makes sense in a variety of contexts (measurable dynamics, piecewise smooth dynamics, etc.) and it models the chemical associa-

[46]The measure on V typically comes as the push-forward of that on X.

tions/dissociations of molecules when V_1 represents the configuration (phase) space of molecules entering the reaction and V_2 describes the product of the reaction, while W corresponds to the short life transition states.

In many biological situations, the phenomenological reduction of the "universal" space X is too drastic to recapture X. The purpose of molecular biology is to find sufficiently many observable parameters needed for reconstructing the point dynamics on X.

Isoperimetry, Cheeger constant and the transition rate. Let X be a space with a fuzzy dynamics A and a distinguished measure μ on X. Define

$$m_A(\mu_0, t) = \mu(A(X_0, t))$$

for $X_0 \subset X$ and $\mu_0 = \mu(X_0)$. The class of functions $m_A = m_A(\mu_0, t)$ is called the *isoperimetric profile* of A, where the essential information is encoded by the Cheeger "constant"

$$Ch(\mu_0, t) = \inf_{X_0 \subset X, \mu(X_0) \mu_0} \frac{m_A(\mu_0, t) - \mu_0}{t \cdot \mu_0}$$

The basic invariant of a stochastic dynamics is the rate of approaching equilibrium (transition rate) which can be expressed in terms of the first eigenvalue λ_1 of the infinitesimal generator of A. In geometry, λ_1 can be bound from below in terms of the Cheeger constant and this idea extends to more general stochastic dynamics, in particular those obtained by reduction of point dynamical systems.

The rate of relaxation (transition) problem relevant to biological systems can be formulated either in the point dynamics framework for a given class of phenomenological observables, or in terms of the stochastic dynamics of a reduced system. In the former case, we start with a linear subspace Φ in the space of all functions (observables) on X and we look for an upper bound on a suitable norm (e.g. the L_2-norm) on Φ in terms of a similar norm on $t^{-1} \cdot (\phi - A(\phi, t))$, for $\phi \in \Phi$, and for the action of A on observables induced by the action on X. If ϕ is a characteristic function of a subset $Y \subset X$, such a bound can be thought of as an isoperimetric inequality, with the ϵ-boundary of Y defined as $\partial_\epsilon(Y) = \bigcup_{0 \leq t \leq \epsilon} A(Y, t) \cap Y$.

In general, there is no non-trivial bound [47] of the measure $\mu(Y)$ by $\epsilon^{-1} \cdot \mu_{\partial_\epsilon(Y)}$, for (infinitesimally) small ϵ, but such bounds are expected for many

[47] Such bounds are possible for non-amenable measurable actions: thy are always present for actions of Kazhdan's T-groups.

particular observables, especially those lifted from sufficiently "strong" and "regular" factorizations V of X. Here, "strong" means that the fibers of the map f are sufficiently large, e.g. they have a relatively low codimension and they are predominantly transversal to the orbits, while "regular" may refer to the regularity of the map f as well as to the invariants of f under some symmetries of the space X.

There is a variety of techniques in geometry for proving isoperimetric inequalities, where many of them are based on the variational techniques reducing the general problem to the "worst" case in a given class of observables (or domains $Y \subset X$) and then showing that the worst case is not so bad after all. In science, in order to bound from below the rate of transition one looks for an explicit realization of the intermediate states allowing fast transitions.

Example: hybridization. We mentioned earlier in this section how hybridization proceeds in steps via the zipping mechanism ensuring a high transition rate between fully dissociated and fully hybridized states.

Example: regulatory proteins. The association time for a regulatory protein finding the regulatory region on DNA is reduced by a certain affinity of this protein with DNA all along the strand. The association process is divided in two stages: first, the protein finds DNA by randomly moving in the solution. The resulting *weak* binding of the protein to the DNA takes significantly less time than finding the regulatory region on random, since the size of DNA is by far larger than that of the regulatory region. Next, the protein starts a random dance in the vicinity of DNA and eventually finds the regulatory region. This random dance on DNA is more time efficient than a random walk in the solution since the size of DNA is small compared to the volume of the cell.

We have also seen intermediate energy scaffolding in the self-assembly processes as well as in the binding of antibodies to antigens. The scaffolding can be justified by a simple computation showing the drastic reduction of the association time within the range of parameters actually observed in the cell, where the relevant parameters in the case of regulatory proteins are:

- the energies of the weak and strong (at the regulatory region) bindings, the thermal energy and the related rate of diffusion (the random walk in solution);

- the volume of the cell, the size of DNA and of the size of the regulatory region, all of them measured in entropic terms.

The above can be interpreted in terms of *parametric quantitative ergodic theory*, where one is concerned with the rate of convergence $R_{\ltimes}$ of the time average of $\phi \in \Phi$ over the interval $[0, t]$, where the space Φ as well as the dynamics A depend on auxiliary parameters, say τ and θ. For example, τ may signify the magnifying power of an observational device, and the dependence of A on θ may encode the ratio between the energies which goes to infinity for $\theta \to \infty$. The problem consists in evaluating $R_{\ltimes}$ for the parameters $t, \tau, \theta \to \infty$ in a certain coherent way, where the dynamical system and the space of observables may degenerate in the limit.

Such an ergodic theory, incorporating the variational isoperimetric techniques, may prove helpful for finding transition states (scaffolding) enhancing relaxation in cells (e.g. protein folding), and for suggesting numerical algorithms modeling the relaxation process. A more ambitious goal concerns the fundamental transitions from non-life to life, where one may start with developing evolutionary realistic scenarios of replication of dynamical systems.

Population states and population dynamics. In many physical, chemical and biological situations, the configuration (phase) space X is (partially) shared by identical or closely related representatives of several species. For example, an ideal monoatomic gas can be represented by a collection of points in $\mathbb{R}^3$, where atoms do not interact with each other. More generally, a gas is constituted by several species of molecules with internal (classical or quantum) degrees of freedom and with mutual hard-core interactions. The system becomes more interesting if we allow chemical reactions leading to the global configuration (phase) space of fluctuating dimension. In biology, one speaks of populations of bacteria in solution where each bacterium in itself makes a statistical dynamical system.

The (classical) symmetry of a population under permutations of individuals within a given species, can be encompassed by the language of *population states.* A population state over X is a measure π on X where $\pi(Y)$, $Y \subset X$, is interpreted as the number of representatives of a given species in Y. If we deal with several species, this π should be vector valued, where the dimension of the range of this measure equals the number of species. Furthermore, if the species have internal degrees of freedom, the measure may take values in a more sophisticated category than in $\mathbb{R}_+$ or $\mathbb{R}^n_+$. Sometimes, this can be viewed as a measure of some extension of X. For example, for a diatomic gas in $\mathbb{R}^3$, the relevant π is the measure on the space of line elements in $\mathbb{R}^3$.

The above discussion can be formally shifted from X to the space $\Pi(X)$ of populations (measures) $\pi(X)$, where one should be aware of the following issues:

- measures π representing actual populations are of rather special nature. For example one can limit to integer valued measures or to measures with a quite small support in X.

- relevant dynamics on $\Pi(X)$ are not linear, but often the non-linearity is well localized. For example, if one has an ensemble of identical particles with hard-core interactions, then the non-linearity occurs only at the collisions.

- a pertinent definition of fuzzy populations and stochastic populations must incorporate the specificity of the space $\Pi(X)$. The classical example is the Poisson distribution σ on a space X, which is, in the present language, a stochastic population of points (states) in X. This is a probability measure on the space $\Pi = \Pi_{\mathbb{Z}_+}(X)$ of integer valued measures on X, which satisfies the Poisson independence axiom: let Y_i, with $i \in I$, be open subsets in X and let m_i be non-negative integers. Denote by $P_i \in \Pi$ the subset of measures π such that $\pi(Y_i) = m_i$. If the subsets Y_i are mutually disjoint, then

$$\sigma(\bigcap_{i\in I} \Pi_i) = \prod_{i\in I} \sigma(\Pi_i).$$

An essentially equivalent way to express the same idea is by saying that σ defines a *Boltzmann sheaf*, that is a *sheaf* of measurable spaces over X. (The notion of such a sheaf makes sense since the category of measure spaces admits fiber products.)

The formalism of stochastic population may look rather farfetched but it seems necessary for a consistent statistical description of biological systems where one needs an even more refined language describing the local/global features of the system and the rate of decay of correlations between individuals in a population. We shall postpone a detailed discussion of this until another occasion.

Genotype, phenotype and syntactic envelops.

The idea of *genotype* can be formalized in a variety of ways starting from the space $\mathcal{S}$ of 4 letter sequences, for simplicity of a fixed (large) length N. Most naively, a genotype is represented by a single sequence $S \in \mathcal{S}$; next, it may be a subset in $\mathcal{S}$; more appropriately, one deals with probability measures on $\mathcal{S}$, interpreted either as random genomes, or as populations. Speaking of populations, one enlarges the setting by allowing sets of populations and random populations.

It is less clear what are mathematical objects corresponding to the notion of *phenotype*. Intuitively, the "mathematical phenotype" should represent an equivalence class of dynamical systems of the kind seen in the previous section, capturing essential features of their behavior. Eventually, a phenotype reduces to a point in some space of observables or to a probability measure on such a space. Furthermore, a phenotype usually appears not as an individual object but rather as a category of these, where the morphisms correspond to transformations (reductions) of observable quantities.

After having chosen suitable definitions for spaces $\mathcal{G}$ of genotypes and $\mathcal{P}$ of phenotypes, one studies the correspondence (is it a map?) from $\mathcal{G}$ to $\mathcal{P}$.

The first issue is finding a *simple syntactic* description of a given genotype $G \in \mathcal{G}$ with a prescribed (phenotypical) error bound. We think of this description as an enlargement of a given sequence or of a sample of sequences, and call it a *syntactic ϵ-envelop* of G, where "syntactic" refers to a chosen formal language and "ϵ" is the size of the phenotypical error.

As a matter of example, let us briefly indicate syntactic possibilities of describing probability measures on $\mathcal{S}$. The simplest class of measures is given by the product measures with weights assigned to the four letters. Such a description is very different from explicitly writing down an individual random sequence with the distribution law described by the measure. The latter, for a large N, is much longer than the formal description of the underlying product measure: the product measure on the space $\{A, T, C, G\}^{3\cdot 10^9}$ with four equal weights for occurrences of the letters, is essentially described in the above line and a half; a description of a point in this space, e.g. the genome of a human individual, takes at least $10^7 - 10^8$ such lines. On the other hand, the phenotypical effects of two quite different random strings with identical underlying probability measure will be most similar if not identical: no living system run by such a genotype exists or can exists.

Fibered measures. Next class of measures on the space $\mathcal{S} = \mathcal{S}_N$ of words of length N is constituted by *fibered measures* defined by four sequences of functions $p_i : \mathcal{S}_j \to [0, 1]$, where $i = 1, \ldots, 4$, $j = 1 \ldots N - 1$ and $\sum p_i = 1$. Given the functions p_i, one obtains fibered measures $\mu_j(\mathcal{S}_j)$ where the corresponding random sequences are defined by the following rule: the j-th letter in a word is chosen on random with the probability weight equal to the value of p_i evaluated on the previous $j-1$ letters of the word. The complexity of such a measure is understood as the complexity of the functions p_i, where the many options for measuring complexity are offered by the traditional computational complexity theory. (There are many variations of this class of measures, essentially due to different ways (or precisions) of ordering the letters, dividing them into blocks etc.)

Parametric measures. Given a measure μ_0 on $\mathcal{S}_M$ and a *parameterization map* $\psi : \mathcal{S}_M \to \mathcal{S}_N$, we have the push-forward measure $\mu = \psi_*(\mu_0)$ on $\mathcal{S}_N$. The complexity of μ can be measured by the sum of the complexities of μ_0 and of the map ψ, where the latter is understood as for a boolean function. (One could embrace fibered and parametric measures in a single more general definition. Also, one could generalize parametric measures by allowing random maps from $\mathcal{S}_M$ to $\mathcal{S}_N$, that are linear maps between the spaces of probability measures of the corresponding sequence spaces.)

The upshot of the two above constructive definitions is the possibility to introduce a notion of complexity (in fact several such notions) for measures on $\mathcal{S}_N$.

Questions. Given a measure μ_0 (e.g. coming from experiments), a phenomenological map $f : \mathcal{S}_N \to \mathcal{P}$, e.g. for $\mathcal{P} = \mathbb{R}^k$, and an $\epsilon > 0$, when and how can one evaluate (from above and from below) the complexity of a measure μ such that $dist(f_*(\mu), f_*(\mu_0)) \leq \epsilon$, for a suitable notion of distance on the space of measures on $\mathcal{P}$ [48]. (If $\mathcal{P} = \mathbb{R}^k$, then one can use $\parallel \mu(f) - \mu_0(f) \parallel$ for example.)

What happens to complexity of measures and approximations, when $\mathcal{S}_N$ converges to the Cantor set (of infinite sequences), for $N \to \infty$?

Both questions have relativized counterparts, where the complexity is measured modulo additional information. For example, $3 \cdot 10^9$ letters making the individual human genome can be reduced to mere $10^5 - 10^6$ symbols describing single nucleotide variations (SNP's) provided the genome is known for another human.

Remark. The above suggests a (possibly useful) interpolation between

[48] Mathematical statistics is concerned with finding rather explicitly described measures μ, such as Gaussian measures, approximating an "experimental" measure.

two foundations of the probability theory: the Kolmogorov complexity and the traditional measure theoretic approach (also due to Kolmogorov).

Remark. The phenotypical map f is not explicitly given in any realistic situation, and can be treated, to a certain extent, as the above "experimental measure" μ_0 [49].

8 Bibliography

INTRODUCTORY AND POPULAR BOOKS:

Ph. Ball. *Designing the Molecular World.* Princeton Science Library, Princeton University Press. 1994.

D.P. Clark, L.D. Russell. *Molecular Biology Made Simple and Fun.* Cache River Press. 1997.

M. Eigen. *Steps towards life.* Oxford University Press, 1996.

D.S. Goodsell. *The machinery of life.* Copernicus Books, 1998.

A.Yu. Grosberg, A.R. Khokhlov. *Giant Molecules, Here , There and Everywhere....* Academic Press. 1997.

J. Maynard Smith, E. Szathmáry. *The Major Transitions in Evolution.* Oxford University Press. 1997.

BOOKS ON BIOLOGY WITH A MATHEMATICAL TILT AND MATHEMATICS ORIENTED TOWARDS BIOLOGY:

R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press. Reprinted. 2000.

J.H. Gillespie. *The causes of Molecular Evolution.* Oxford Series in Ecology and Evolution, Oxford University Press. 1991.

[49] One can speculate that this map could be derived from fundamental physical laws if we had had an infinite computational power at our disposal. Since this is unavailable, one searches for particular patterns and regularities in specific maps and measures coming from (biological) experiments and allowing a feasible description of these objects.

S.A. Kauffman. *The Origin of Order. Self-organization and selection in evolution.* Oxford UNiversity Press, 1993.

Wen-Hsiung Li. *Molecular Evolution.* Sinauer Associates Inc. 1997.

P.A. Pevzner. *Computational Molecular Biology: an algorithmic approach.* MIT Press, Computational Molecular Biology Series, 2000.

TEXTBOOKS:

B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson. *Molecular Biology of the Cell.* Garland Publishing Inc., Third Edition. 1994.

V.A. Bloomfield, D.M. Crothers, I. Tinoco Jr. *Nucleic Acids, Structures, Properties, and Functions.* University Science Books, 2000.

R. Chang. *Physical Chemistry for the Chemical and Biological Sciences.* University Science Books, 2000.

H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J.E. Darnell. *Molecular Cell Biology.* W.H.Freeman and Company, Fourth Edition. 2000.

L. Pauling. *General Chemistry.* Dover Publ., 1989.

T. Strachan, A.P. Read. *Human Molecular Genetics 2.* BIOS Scientific Publishers Ltd., Second Edition. 1999.

L. Stryer. *Biochemistry.* 4th edition, W.H. Freeman and Company. 1995.

D. Voet, J.G. Voet. *Biochemistry.* John Wiley and Sons, Second Edition. 1995.

A FEW ARTICLES SUGGESTED TO THE READER:

J.-F. Allemand, A. Bensimon, D. Bensimon, F. Caron, D. Chatenay, Ph. Cluzel, V. Croquette, Ch. Heller, R. Lavery, A. Lebrun, T. Strick, J.-L. Viovy. L'ADN ressort moléculaire. *Pour la Science*, 224:76–82, 1996.

Ph. Cluzel, A. Lebrun, Ch. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, F. Caron. DNA: an extensible molecule. *Science*, 271:792–794, 1996.

E.D. Green. The human genome project, and its impact on the study of human disease. Chapter 9 in the 8th edition of *Metabolic and Molecular Bases of Inherited Disease*, edited by C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle, B. Childs and B. Vogelstein, 2000.

P. Green. Transduction to generate plant form and patterns: an essay on cause and effect. *Annals of Botany*, 78:269–281, 1996.

A. Hénaut, T. Pouxel, A. Gleizes, I. Moszer, A. Danchin. Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *Journal of Molecular Biology*, 257:574–585, 1996.

A. Hershko, A. Ciechanover, A. Varshavsky. The ubiquitin system. Basic Medical Research Award, *Nature Medicine*, 6(10):iii–xi, 2000.

F. Jacob, J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356, 1961.

L. Mendoza, D. Thieffry, E.R. Alvarez-Buylla. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics*, 15:593–606, 1999.

N. Seeman. DNA nanotechnology: from topological control to structural control. In *Pattern Formation in Biology, Vision and Dynamics*, A. Carbone, M. Gromov, P. Prusinkiewicz eds., World Scientific, 2000.

A.A. Simpson et al. Structure of the bacteriophage ϕ29 DNA packaging motor. Letters to *Nature*, 408:745–750, 2000.

P. Smolen, D. Baxter, J.H. Byrne. Mathematical modeling of gene networks. Review in *Neuron*, 26:567–580, 2000.

P. Smolen, D. Baxter, J.H. Byrne. Modeling transcriptional control gene networks - methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62:247–292, 2000.

S. Tyiagi, D.P. Bratu, F.R. Kramer. Multicolor molecular beacons for allele discrimination. *Nature Biotechnology*, 16:49–53, 1998.

A. Varshavsky. The N-end rule pathway of protein degradation. Review Article, *Genes to Cell*, 2:13–28, 1997.

E. Yeramian. The physics of DNA and the annotation of the *Plasmodium falciparum* genome. *Gene*, 255:151–168, 2000.

E. Yeramian. Genes and the physics of the DNA double-helix. *Gene*, 255:139–150, 2000.

C.-H. Yuh, H. Bolouri, E.H. Davidson. Genomic Cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.

Tables and Figures:

Some of these have been adapted from the textbooks cited above, where the reader can find more detailed information.