Crystalls, Proteins and Isoperimetry.

Misha Gromov

September 29, 2009

Contents

Abstract

We attempt to formulate several mathematical problems suggested by structural patterns present in bio-molecular assemblies. Our description of these assemblies, by necessity brief, is self-contained, albeit on a superficial level. If a reader stumbles upon something confusing, this, hopefully, will become clearer at a later point in the article.

1 Is there Mathematics in Biology?

Day-Night, Day-Night, Day-Night. Summer-Winter, Summer-Winter... .

What is the structure behind this. Where should we look to uncover it? Is it in the singing of birds, in the flows of rivers, in the changes of temperature?

We know, with hindsight, what we should do.

Look up into the night sky, find a few specks of light – planets – that slowly crawl amongst the unmoving stars, imagine how their routes appear if seen from the sun, invent calculus, guess differential equations for the planetary routes, unravel symmetries of these.... and the world of mathematical wonders opens to you: Lie groups, algebraic varieties, symplectic manifolds... .

Now try: Alive-Dead, Alive-Dead, Alive-Dead,... . Where do we go from here?

"... living matter, while not eluding the "laws of physics" as established up to date, is likely to involve "other laws of physics" hitherto unknown, which however, once they have been revealed, will form just as integral a part of science as the former".

Erwin Schrdinger, who wrote this in 1944 in his book "What Is Life?", apparently, had in mind some counterpart to the Second Law of Thermodynamics. The "other laws of physics" have not materialized. But biology has gone far since 1944 – the molecular patterns of Life have been displayed before our eyes as the stars in the night sky.

But what are the specs of light that would guide us to the world of new mathematics?

At the first sight, Nature does not appear exceptionally clever, her evolutionary strategy is not sophisticated, to say the least. But she was selecting from billions upon billions of candidates and her selection criterion "fit to survive" may look simple only for a lack of mathematical imagination on our part: enormous amount of structure goes into this "fit". Besides, Nature does not run in structural vacuum: all of physics and chemistry is at her disposal, she excels in molecular dynamics and in catalysis.

Yet, a mathematician might think that Nature is dumb: the primitive mutation/selection mechanism of evolution could not produce anything we, mathematician, could not divine ourselves.

But if so, we inevitably conclude that the human brain, which was cooked up by Nature in the last couple of millions years can not be especially smart either: all our mathematics, or rather the mathematics building mechanisms in the brain, must be confined to the rules that the evolution had stumbled upon in this relatively short stretch of time and had installed into us.

On the other hand, Nature had spent much longer time (measured by the number of tries involved) in inventing such structural entities as the cell and the *ribosome*.

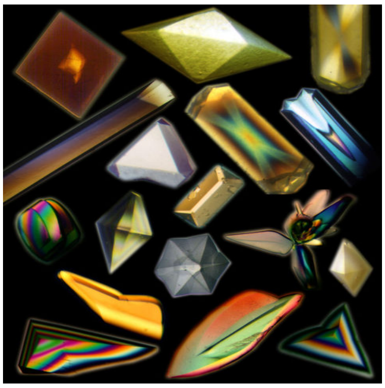
(Ribosomes are large molecular assemblies $\approx 20nm = 2 \cdot 10^{-6}cm$ in diameter composed of *ribosomal* RNA and proteins. As a ribosome crawls along a *messenger* RNA it synthesizes a *polypeptide chain* out of 20 (+1) amino acids in the cell by *translating* genetic information written on this RNA in four letters – four species of basic units – *nucleotide* molecules, where an RNA is a hundreds/thousands long polymer chain composed of these units.

There are usually many ribosomes translating in parallel from a single RNA molecule with ≈ 100 or less nucleotides between them; one might say that it is RNA who "*crawls*" through a train of ribosomes.)

One may *conjecture* that neither cell nor brain would be possible, if not for profound mathematical "somethings" behind these Nature's inventions. But what are these "somethings"? Why do we, mathematicians, remain unaware of them?

Notwithstanding our much glorified successes we are, tautologically, blind to what we do not see. (Nature systematically hides from our mind what we are not supposed to know, such as the blind spot in our retina, for instance. The neurological mechanism of this hiding is far from clear.)

Also, the history of mathematics shows how slow we are when it comes to inventing/recognizing new structures even if they are spread before our eyes, such as the hyperbolic space, for instance. (More recent and more relevant examples are seen in the slow start in mathematical development of Mendelian



genetics and in the failure of identifying general mathematical principles underlying Sturtevant's reconstruction of the linear structure on the set of genes on a chromosome of Drosophila melanogaster from samples of a probability measure on the space of gene linkages [8].)

Our brain is hardly able to generate mathematical concepts by itself, it needs an input of "raw structures" and Nature has much of them to offer. The problem is that this "much" which the biology offers to us is "too much": it is hard to decide what in this offer contains germs of new mathematics and what is a "frozen accident" – an irrelevantly special complexity.

The only way to reject the irrelevant is to first learn and understand what it is. One has to browse through myriads of stars – structural specks of Life revealed by biologists – in order to identify the "essential ones". And when (if?) we find them, we may start on the long road toward new mathematics.

And even if we fail to assemble a coherent structure from the multitude of available fragments, we may gain a better vision on the boundary of our mathematical knowledge which is hidden from us by the "complacency wall" of an intrinsically mathematical point of view.

2 Periodic Molecular Assemblies.

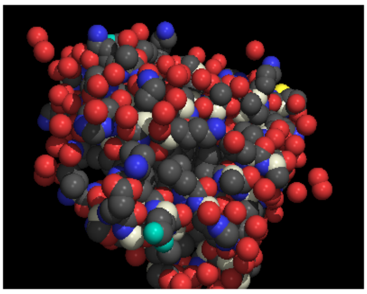
The $\mathbb{Z}^3$-symmetry of crystals, i.e. the triple periodicity on the atomic/molecular scale, was discovered/conjectured by René Just Haüy in the late 1700's who, according to mineralogists' lore, came to the idea while pondering over the fragments of a broken calcite crystal. (His "Traité de Minéralogie" appeared in 1801.)

But why symmetry? It seems unlikely that if you shake potatoes in a big box they would spontaneously arrange themselves into something symmetric; yet, this happens to many large potato-shaped molecules, such as *myoglobin* (pictured below) – a protein with $\approx 3nm$ diameter molecules ($1nm = 10^{-9}m$) made of about 2 000 atoms – 2-3Å-balls (1Å=0.1nm.)

(Myoglobin stores oxygen in muscles; it contains a metal-organic *heme group* with an Iron atom to which O_2 binds.

The 3-D structure of myoglobin was the first protein structure solved by John Kendrew in 1958 with *X-ray diffraction analysis*

The diffraction of X-rays delivers *only the amplitude* of the Fourier transform



of the electron density in a crystal; the periodicity of the crystal is crucial for the extraction of the information on individual molecules on the Å-scale from the diffraction image; this involves many non-trivial mathematical and non-mathematical ideas, see ”X-ray crystallography” in Wikipedia)

How does the symmetry come about? The easiest to account for is the *helical symmetry* of molecular assemblies [5],[27], [12], [14]. Suppose that two molecular (sub)units of the same species M preferentially bind by sticking (docking) one to another in a certain way (i.e. the binding energy for a pair of molecules has a unique minimum sufficiently separated from other local minima). If $M_1 \vDash M_2$ is a pair of so bound molecules in the Euclidean space $\mathbb{R}^3$ then there is a (typically unique) *isometric transformation* (rigid motion) T of $\mathbb{R}^3$ moving M_1 to M_2.

Such a T, by an elementary theorem, is made by a rotation around an axial line L in $\mathbb{R}^3$ followed by a parallel translation along L. If the copies $M_1, M_2 = T(M_1), ..., M_n = T(M_{n-1})$ do no overlap, the chain of n copies of M, written as $M_1 \vDash M_2 \vDash ... \vDash M_n$, makes a helical shape in the space around L which provides a minimum of the binding energy to the ensemble of n molecules. This minimum, even if it is only a local one, has a rather large *attraction basin* (this needs a proof) which make the helical arrangement kinetically quite probable.

Helical symmetries (α-*helices*) are ubiquitous in proteins as was determined by Pauling, Corey, and Branson in 1951 on the basis of the structures of amino acids and the planarity of *peptide* bonds between amino acid (residues) in proteins. (Another commonly present pattern in proteins, called β-*sheet*, displays $\mathbb{Z} \oplus \mathbb{Z}$ symmetry, see section 5.)

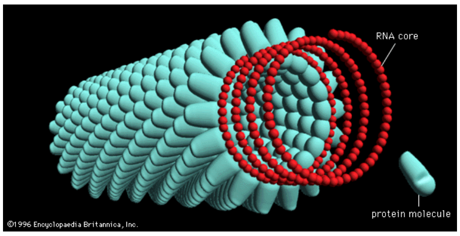
DNA molecules also have helical symmetry. (Three types of DNA double-helices have been found). A helix is composed of two polymer chains held together by *hydrogen bonds.*

(These chains, reaching $\approx 2{\cdot}10^8$ in human chromosomes, are made, like RNA, from four species of basic units – *nucleotides*: **A**denine, **G**uanine, **C**ytosine and **T**hymine which are composed of 15 ± 1 atoms each and which are similar to the molecular units making RNA's.)

Another kind of helix, making a rode-shaped viral particle about 300nm($= 3{\cdot}10^{-5}$cm) long and 15-20nm in diameter, is that of *Tobacco Mosaic Virus capsid* (exterior shell) made from 2130 molecules of coat protein. (Imagine how one gets this ”2130” or look into [24].)

(Dmitri Ivanovski in 1892 provided evidence for a non-bacterial agent affecting tobacco plants which remained infectious after fine filtering.

Wendell Stanley isolated and crystallized the virus in 1935 and showed that



it remains active after crystallization.

Gustav Kausche, Edgar Pfankuch and Helmut Ruska obtained the electron microscopical images of the virus in 1939.

James Watson studied the X-ray diffraction on the virus crystals in 1952-1954 and deduced its helical structure.

Heinz Fraenkel-Conrat and Robley Williams showed in 1955 that purified viral RNA and its capsid protein assemble by themselves to functional viruses.)

The "biological helices", are "more symmetric" than the "naked helix" – (almost) every subunit has *three or more* neighbors bound to it. The origin of "bio-helical" and non-helical symmetries can be understood by looking at the 2-dimensional version of the rigid motion T in the 3-space.

A typical isometry T of the plane is a rotation by some angle α around a fixed point, where this α is determined by the "binding angle" of $\vDash$ between M_1 and M_2. If $\alpha = 2\pi/n$ where n is an *integer*, then the n copies of M make a roughly circular shape with the n-fold rotational symmetry.

Thus, for example, not all molecular units M could form a 5-fold symmetric assembly, but if the "$\vDash$-angle" is (planar and) close to $2\pi/5$ and if the $\vDash$-bond is slightly flexible then such assembly, with every copy of M involved into *two* "slightly bent" $\vDash$-bonds, will be possible. (Honestly, I do not know what is an appropriate description of "bending" of a *quantum*-chemical bond on the relevant energy/space scale.)

Returning to the Euclidean 3-space $\mathbb{R}^3$, if one wants a sufficiently rigid molecular assembly (e.g a viral shell) V, where there are more than two, say 4 neighbors for each copy of M with two different kinds of $\vDash$ bonds and such that V admits say, two symmetries T and T', then one needs to satisfy certain relations between the "$\vDash$-angles" of mutually bound molecules, similar to but more complicated than the $2\pi/n$-condition.

These geometric relations must ensure the algebraic relations between the generators T and T' in the expected symmetry group Γ, where Γ is not given beforehand – it comes along with the self assembly process and may depend on specific kinetics. For example, the symmetry of protein crystals, one of 230 possible crystallographic groups, may depend on the particular condition at which a given protein is being crystallized (see "crystal structure" in Wikipedia and [10].)

The isometry group Γ itself does not determine the geometry of a Γ-symmetric assembly: specific generators T, T',...of Γ are essential. For example, the helical symmetry is governed by the group $\Gamma = \{..., T^{-2}, T^{-1}, T^0 = \mathbf{1}, T^1, T^2, ...\}$, which is isomorphic to $\mathbb{Z}$ – the additive group of integers. This $\mathbb{Z}$, in "biological helixes", is given by two or more generators, say, by T and T' with the relation $T' = T^n$ with a moderately large n, e.g. $n = 4$ for α-helices, where T

is associated with the (strong covalent) *peptide* bonds and T' with (weak) *hydrogen bonds* between amino acid residues in a protein. (Sometimes $n = 3$ and rarely $n = 5$ in the α-helices.) The Tobacco mosaic virus capsid has $T' = T^n$ for $n \approx 16.3$. (Non-integer?! – Does the virus defy math? Are the T'-bonds weaker than the T-bonds? Are they non-specific?)

Alternatively, let us think in terms of the full configuration space $\mathcal{M}$ of molecules M, where there is an action A, or a family of actions, by a group Γ on $\mathcal{M}$. If a (energy) function E on $\mathcal{M}$ is invariant under these actions, then one can easily show in many cases that the local minima of E on the *subspace* of Γ-*symmetric* configurations also serve as local minima on *all* of $\mathcal{M}$.

For example, $\mathbb{R}^3$ admits a 9-parameter family of actions A by the group $\Gamma = \mathbb{Z}^3 = \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}$ (where each A is generated by 3 parallel translations) which induce in an obvious way actions A of Γ on the (infinite dimensional) configuration space $\mathcal{M}$ of molecules in $\mathbb{R}^3$. Since the (mean) binding energy between molecules is invariant under A, the appearance of organic (tri-periodic) crystals (regarded as minima points in the configuration space of molecules) looks less miraculous.

(The original group acting on the space $\mathbb{R}^3$ is its full isometry group, where a specific discrete subgroup Γ comes in the process of a particular molecular assembly.)

Crystals and crystal growth have been much studied by mathematical physicists but I doubt that a *comprehensive purely mathematical* model incorporating an evaluation of the attraction basin of a crystal and a rigorous quantitative kinetics of crystallization is available at the present day. Here are questions.

How much does a symmetry of a molecule M with a finite group G which lies in the rotational quotient of some crystallographic group Γ enhance the attraction basin of Γ-crystals of M?

How does a mixture of several *different kinds* of molecules crystalize?

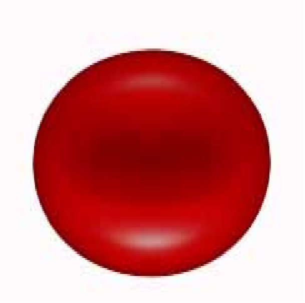
For instance, protein crystals, retain about 30% of water molecules which play an essential role in the crystal formation [18].

One may think (on the basis of counting parameters) that a mixture, say of M, M' and M'' with a random proportion $c : c' : c''$ of the concentrations would not crystallize but there are particular $c : c' : c''$ highly beneficial for crystallization, especially if M' and M'' are smaller than M and nicely fit into the gaps between the M-molecules. (I guess, everything of this kind is known to crystallographers [17] but, in the long run, mathematicians may come with something practically useful.)

And forgetting "physical crystals", a mathematician may wonder if something like discrete subgroups of adelic groups also come by a process of crystallization from some huge "configuration space".

3 Crystals in Fluids: Erythrocytes, Liposomes, Micelles.

An erythrocyte – a red blood cell – a carrier of *haemoglobin* – is a roughly rotationally symmetric cell, kind of a biconcave disk (thin near the center and



the thickest at the periphery) of diameter $6000-8000nm$ and of thickness $1500-1800nm$ (small by the animal cells standards.) The membrane (surface) of an erythrocyte is a 6-8nm thick *bi-layer* of rod-like (phospholipid) molecules oriented normally to the surface of the membrane with *hydrophilic* "heads" facing the exterior and the interior of a cell while the *hydrophobic* "tails" are buried inside the membrane. Such construction behaves as an incredible (quasi-inertionless) 2D-fluid curved in 3-space: free to move within itself (preserving area) but resisting bending.

Erythrocytes' (idealized) shape is believed to be a solution of an isoperimetric problem being a (closed simply connected) surface S in 3-space, with prescribed both, area and 3-volume it bounds in space, that *minimizes* the integrated squared curvature encoding the bending energy of S. (See [11] and "Elasticity of cell membranes" in Wikipedia)

The corresponding variational/isoperimetric problem is easy in the class of rotationally symmetric surfaces but I doubt that the rotational symmetry of the extremal surface has been proved.

(A similar picture is seen in the spherically symmetric shapes of small drops of liquids, of *micelles.* and of *liposomes.* Spheres "solve" the *isoperimetric problem*: they surround given volume by a surface of minimal area.

This is attributed to *Dido* who, according to ancient Greek and Roman sources, had solved the $2D$-isoperimetric problem in the course of founding her kingdom of Carthage in ≈ -900. But some historians are doubtful that Dido was influenced by erythrocytes, that she realized that not all symmetric problems necessarily had equally symmetric solutions and even that she could furnish a rigorous proof of the sphericity of the minimizer to $area/volume^{\frac{2}{3}}$.)

The above motivates the following geometric variational problem.

Let X be C^∞-smooth n-manifold, let an integer k be given, let X' denote the Grassmann bundle of tangent k-planes in X and let $X^{(r)}$ be defined by $X^{(r)} = (X^{(r-1)})'$.

If X carries a C^∞ Riemanninan metric g_X than g_X and the $O(n)$-invariant metric g_{Gr} in the Grassmanian fiber $Gr_k(\mathbb{R}^n)$ define, via the Levi-Cevita splitting of the tangent bundle $T(X')$, a family of metrics $g'_{p_0,p_1} = p_0 g_X + p_1 g_{Gr}$ on X', for all $p_0, p_1 > 0$.

Similarly, $X^{(r)}$ carries a family of metrics g^r_p parametrized by the positive cone $P = \mathbb{R}^{r+1}_+ \ni p$.

Every smooth k-submanifold $S \subset X$ lifts to $S^{(r)} \subset X^{(r)}$ and the k-volumes of these lifts with respect to g^r_p define a function $vol^r : \mathcal{S} \times P \to \mathbb{R}_+$, where $\mathcal{S}$

denotes the space of all S and where vol^r is regarded as the family of functions vol^r_p on $\mathcal{S} = \mathcal{S} \times p$.

What are the critical points of vol^r_p?

These only rarely belong to $\mathcal{S}$ itself, one needs to properly complete $\mathcal{S}$ (with respect to a suitable metric or otherwise) in order to allow certain singularities.

What is the (minimal) completion(s) for this? What are the singularities of extremal S?

Prominent examples of possibly singular vol^r_p-minimizers are complex subvarieties S in algebraic/Kähler manifolds X. Probably (this seems easy) these S, even if singular, are stable as critical points under small non-Kählerian perturbations of the function vol^r_p (with suitably understood "stability" for non-isolated S).

How much does the function vol^r_p vary with p on the subset of extremal S? In particular, how does it blow up (if at all) at the boundary of P?

How much does the picture change if we use another metric on $X^{(i+1)}$, e.g. associated to non-Levi-Civita connections on the bundles $X^{(i+1)} \to X^{(i)}$ for splittings the tangent bundles of $X^{(i+1)}$.

What is the symmetry of the solution(s) to the corresponding Erythrocytes' isoperimetric vol^r-problem for hypersurfaces in $\mathbb{R}^n$ for given r and $n >> r$.

What happens if, instead of the tangent k-planes, one uses tangent l-planes to S for $l \neq k = dimS$, e.g. for $l = n - 1 = dimX - 1$ on the first step?

The mathematics becomes more involved if we recall the physical origin of our surfaces S which serve as boundaries of *liposomes* and, similarly, of *micelles* – interfaces between water and hydrophobic substances (see "lipid bilayer" in Wikipedia). These emerge as (low temperature) limits of statistical ensembles $\mathcal{SE}$ (of many particle systems) of asymptotically infinite dimension and the geometric PDE satisfied by such S, (e.g. minimality, constant mean curvature, etc.) can be derived from the corresponding geometric properties of $\mathcal{SE}$.

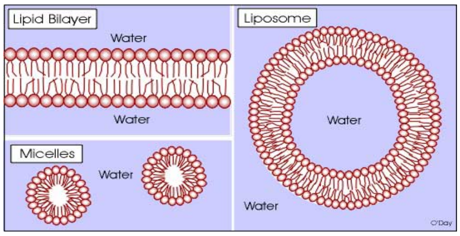
(It is tempting to model such ensembles in the algebraic/Kählerian case by limits of probability measures on families of algebraic subvarieties $S^\perp$ in X of dimension $2m = dimX - dimS$ and degree $D \to \infty$ which are eventually transversal/normal to S and which play the role of bylipid rode-shaped molecules; alternatively, one may use similar subvarieties in $X^{(r)}$ transversal/normal to the lift $S^{(r)}$ of S to $X^{(r)}$, or else, closed positive (m,m)-currents in X transversal/normal to S.)

Can one go the other way around and find a (quasi)functorial gate (arrow) from some classes of partial differential equtions $\mathcal{D}$ (and/or geometric structures underlying $\mathcal{D}$) to (asymptotically infinite dimensional) statistical ensembles $\mathcal{SE}$. Can one gain in symmetry while passing through such a gate? Is there some asymptotic expansion of (stochastic) perturbations of an extremal S in the space $\mathcal{SE}$?

(Possibly, something appears in physics literature but I could not trace anything in mathematics. A topological, rather than statistical "gate" of this kind is suggested in [7] but it has a rather limited range of applications.)

What do micelles/liposomes have to do with crystals?

Both can be regarded as low temperature limits of statistical ensembles of particle systems with short range interactions (this is an oversimplification as we shall see below) where in many (but not all) cases a single "particle" M can be reperesented by an (tangent) orthonormal frame in space with the total energy



of the ensemble being $\sum_{ij} E(M_i, M_j)$ for some binary (pairwise) interaction (potential) energy function $E : Frame(\mathbb{R}^3) \times Frame(\mathbb{R}^3) \to \mathbb{R}$ (with the usual contribution of the kinetic energy and the entropy terms).

Also, "packings" of S by "rods" are similar to packings of molecules (in $\mathbb{R}^3$) into crystals. Probably, a "packing of an extremal surface S by rods" is combinatorially periodic on S minus a "small subset" $\Sigma \subset S$ (most likely, $codim\Sigma = 2$), provided the minimal packing of the plane by the cross-sections of the rods is stably periodic.

Finally, assume the "rods" are cylindrical and replace them by balls with the the same cross-sections. The arrangement (or rather a small rearrangement) of these balls along S, originally made by the rods, remains extremal but does not minimizes the energy anymore (due to the excess of positive curvature of the spheres which bound the balls): the Morse index of the corresponding critical point of the energy will be around N – the number of rods (or balls) going into making S.

This brings "liposomal" S on equal footing with crystals as far as the topology is concerned: such surfaces S are associated with certain extremal (not minimal as for crystals) points of the energy on the configuration space of (infinitesimally small) balls. This agrees with the (geometrically) natural (co)homological coupling between spaces of $\mathbb{Z}_2$-cycles in Riemannian manifolds X and configuration spaces of balls in X studied in [9], but it remains unclear how measure, entropy and topology blend together.

Is there a meaningful picture for extremal arrangements of "rods" along S, where the "essential dimension" of these "rods" is *strictly* between 0 (as it is the case for balls in [9]) and $codimS$ (as in [7])?

Are there stochastic extensions/perturbations of the discrete groups in non-Euclidean Lie groups similar to these for micelles and liposomes?

Do subvarieties of algebraic varieties over fields of finite characteristics admit "liposome-crystal" models?

Water, Hydrogen Bonds and Hydrophobicity. Biological molecules live in water and interact (e.g. attract one to another) via *weak* chemical bonds.

The strength of a chemical bond (interaction) is measured by the energy needed to break it, where a convenient reference energy is Boltzmann's $\frac{3}{2}kT$ (where 3 is the dimension of the space and $\frac{1}{2}$ comes from $E = \frac{mv^2}{2}$) at the room temperature $T = 298K \approx 25°C$, that is the average kinetic energy $\frac{mv^2}{2}$ of molecules in a liquid (or gas) which is the same for all molecules regardless of their mass m. For example, the (square) average speed v of water molecules is about $650m/s$ at the room temperature.

The room temperature kT is close to 2.45 kJ/mol ≈ 0.6 kcal/mole and to

1/40 eV in the standard absolute units (1kcal/mol $\approx$ 4.1840 kJ/mol and eV$\approx$ 96.5 kJ/mol, where $1/mol = N_A^{-1}$ and $N_A \approx 6.0221415 \times 10^{23}$ is the *Avogadro number* – the number of atoms in $12g$ of carbon ^{12}C by the presently accepted normalization.)

For comparison, a green photon carries about $2.5eV \approx 100kT$ of energy and most covalent bonds have comparable energy: they are stable under the room temperature if not exposed to light. Peptide bonds in this sense are unstable in the presence of water. In fact they *release* about $10kJ/mol \approx 4kT$ when they break (by *hydrolyzing*, i.e. taking back water molecules which they loose when a protein is synthesized). But an easy break at the room temperature is prevented by an "energy barrier": the half-time for hydrolysis under physiological conditions is a few hundred years [21] [22]. (Also the *phosphodiester* bonds between nucleotides monomers in DNA and RNA are *meta*stable in water.)

Weak chemical bonds in biological molecules are somewhere within 1-6kT (sometimes more) where the strongest among the weak are the so called *H(ydrogen) bonds* which are due to electrostatic attraction between molecules with non-uniformly distributed charges associated with (displacements of) protons of particular hydrogen constituents of molecules.

Bonds close to kT are not static: they constantly break and reappear in thermal equilibrium by exchanging their energy with the molecular kinetic energy; also the effective H-bonds between biological molecules M (e.g. amino acids in proteins) are weakened by exchange with H-bonding between M and water molecules.

The H-bonds between water molecules themselves (3-5kT) make a complicated dynamic network (the properties of which is still not fully understood on the nano-scale) which makes the thermodynamics of water quite peculiar. For example the boiling temperature, of water (H_2O of atomic weight ≈ 18) is quite high ($100°C \approx 373K$) compared to other substances of comparable and even larger molecular weight (e.g. $\approx 90K \approx -183°C$ for O_2 of weight ≈ 32 and $\approx 216K \approx -57°C$ for CO_2 of atomic weigh ≈ 44).

Collectively, weak bonds may be quite stable, as in *folded proteins* at the room temperature for instance, but the *pronounced energy gap* between weak and covalent bonds seems essential for the function of biomolecules,

(Do not think kT is too small to bother about: if the weak interactions in your proteins went down by 2%, which would amount to the raise of the body temperature by $\approx 6K$, you, above $43°C \approx 316K$, would be as good as dead.

Yet some thermophilic unicellular organisms, most of them are *archaea*, strive above the point where all your proteins would unfold. For example, the incredible *Strain 121*, isolated from a thermal vent deep in the Pacific Ocean, reproduces at $121°C$ and survives for several hours at $130°C$.

But *Thermus aquaticus*, whose *DNA-Polymerase* is used in commercial PCR for the diagnostic DNA amplification, is not an archaeon – it is a thermophilic *bacterium*.)

Hydrophobic molecules, such as (phospho)lipids are not (significantly) polarized and form no hydrogen bonds with water. However, the presence of these molecules in water disrupt hydrogen bonds between water molecules; thus, the whole system "tries" to minimize the interface between water and a hydrophobic substance.

The resulting surfaces S, boundaries of micelles for instance, *are not* locally minimizing, however, in the class of *smooth surfaces*, since the water molecules on different sides of S do not (significantly) interact across S. The (local) minimality is manifested in *discontinuous* stochastic perturbations of S in $\mathcal{SE}$. The formation of such an S in $\mathcal{SE}$ does not follow a simple energy gradient curve. It is rather a "gradient tree" something like a branched network of tributaries of (the bed of) a river along which the energy flows downhill, and where the "branches" are physically implemented by disconnected surfaces with boundaries playing the role of *nucleating sites* ("seeds") in the standard picture of *crystal growth* (see Wikipedia).

All this points to possible stochastic extensions/pertuprbations of the vol^r-model of liposomes and micelles but mathematics of this is nonexistent yet.

Erythrocytes and Haemoglobin. Erythrocytes carry *haemoglobin* in the blood of an animal body where they manage to squeeze through $5000nm$ thin capillaries without major distortion. (Yet, high pH, high calcium concentrations, exposure to glass surfaces, reduced albumin concentrations, and prolonged storage turn erythrocytes into *crenated*, also called *burr cells* with short, sharp spikes.)

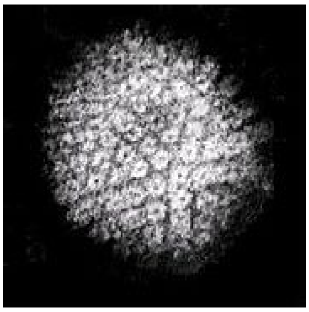
Haemoglobin is a large (roughly spherical $\approx 6nm$ in diameter) protein built of four structurally similar non-covalently (weakly) bound subunits totally of 574 amino acids (residues). Each subunit contains a set of alpha-helix segments spatially arranged in a particular *globin* (supersecondary) pattern, which incorporates a *heme group* that is an organic molecule with an iron atom in it. This is instrumental for binding Oxygen in the lungs and transporting the bound oxygen throughout the body and carrying in exchange CO_2 from tissues back to lungs.

The oxygen released from hemoglobin in muscles binds to myoglobin which stores oxygen in muscles (and also contains for this a heme groups with an Iron atom).

The binding of Oxygen to hemoglobin is a positive cooperative process: when one subunit in haemoglobin becomes oxygenated it induces a *conformation change* in the whole protein causing the other three subunits to gain in affinity for oxygen. (This is a common Nature's trick: use k copies of "something" in order to sharpen an s-response to the collective threshold-like s^k.) Thus, haemoglobin (unlike the single unite myoglobin) switches from binding oxygen in lungs to its release in muscles where the partial O_2-pressure drops ≈two-fold. (This "two" is roughly, the same as the ratio between the see level pressure and that at 6 km, from where up breathing becomes a problem for humans. Yet, certain birds, e.g. some geese and vultures, fly comfortably above 10 km with one quarter of the see level pressure.)

Haemoglobin makes about 97% of the red blood cells dry content and needs no complicated chemical purification. It was the first protein to be crystallized. The crystals were obtained by Otto Funke (and Karl Reichert?) around 1850 by diluting red blood cells with a solvent followed by slow evaporation. (Most pure proteins crystallize only under particular special conditions and/or after a modification of the molecules. The symmetry of the Haemoglobin molecule, probably, facilitates crystallization.)

Erythrocytes are continuously produced in the red bone marrow of large bones (in adult humans at the rate≈ 2.5 million/sec. or ≈ 200 billion/day) and in the mature form (in mammals) contain no DNA and do not synthesize



their proteins. Adult humans have 20-30 trillion erythrocytes, ≈ 5 million per cubic millimeter of blood; a human erythrocyte contains about 200-300 million hemoglobin molecules.

All this is just a speck of foam in the sea of biological knowledge. Where does the structure starts and where does it end in this sea?

4 Information and Symmetry in Viruses.

In 1956, Crick, Watson, Caspar and Klug had *predicted* possible *icosahedral symmetry of viruses* (such as *Herpes simplex virus* in the above TEM micrograph) by an essentially *mathematical* reasoning partly based on analysis of X-rays diffraction on crystals of viruses [26] [27].

Why does the *random* mutation/selection evolutionary mechanism generates *improbable* symmetries?

The underlying physical reason for this must be apparent by now: the symmetries of viruses, similarly to crystals, reflect the spacial symmetry of the physical laws as it was pointed out in [3] p. 3.

"Self assembly (of a virus) is a process akin to crystallization and is governed by the laws of statistical mechanics. The protein subunits and the nucleic acid chain spontaneously come together to form a simple virus particle because this is their lowest (free) energy state".

The symmetry of viral capsids which has been well established for many virusus (e.g. the viruses of the herpes family are icosahedral ≈100nm in diameter) [2] [25], is no more paradoxical than that of a protein crystal. In a nutshell, if one is ready to disregard the uncomfortable fact that symmetric equations may have non-symmetric solutions, one might simply say that since the physical world is symmetric, symmetric forms are likely to be functionally as good, if not even better, than non-symmetric ones. For example, the bilateral symmetry of our bodies is good for walking. (Symmetry is persistent in linear systems, e.g. in small oscillation of viral capsids [1], [23].)

The above geometric/physical consideration shows that viral symmetry is plausible but not necessarily *very* probable. The decisive reason for the symmetries of viral shells (capsids) set forth by Crick and Watson was that a virus needed to pack "maximum of genetic information" in a small shell (capsid) which is built of proteins encoded by the viral genes.

(The idea that DNA codes for proteins was in the air since the 1953 reconstruction of the DNA double helix structure by Crick and Watson, partly based on the X-ray diffraction results of Franklin and Wilkins. Gamov – the author of the big bang theory – suggested in 1954 that each of 20 amino acids must be coded by a triplet of nucleotides, since $n = 3$ is the minimal solution to the inequality $4^n \geq 20$, where 4 is the number of different species of nucleotides in DNA. Gamov's is an amazing instance of a simple idea in biology which turned out to be true.)

Indeed, if a virus uses n genes in its DNA (or RNA) for the shell proteins, say, each codes for m copies of identical proteins molecules, then the resulting viral shell can contain DNA of size $\sim (nm)^{\frac{3}{2}}$. If m is large, this allows smaller n which is advantages to the virus. (A small virus replicates faster since more copies of it are yielded by an invaded cell.)

The above energy argument implies that the presence of equal copies of protein molecules make symmetric assembles quite likely if one properly adjust the ”$\vDash$-angles”

Now, an evolutionary factor enters the game: a symmetric form can be specified by fewer parameters than a functionally comparable non-symmetric one. For example, a non-symmetric assembly of molecules may have many different ”$\vDash$-angles” all of which need be somehow encoded by the viral DNA (or RNA) while a symmetric form has many of these ”angles” mutually equal. This simplifies Nature's task since it selects from a smaller pool of competing possibilities.

(The presence of *many identical* copies of *large heterogeneous* ”units”, e.g. heteropolymeric molecules, is the hallmark of life. These are produced by some *universal* processes of *controlled amplification* – another characteristic feature of living systems. The basic instances of this are *transcribing* (by ”templating”) messenger RNA from DNA and then translation from messenger RNA to proteins by ribosomes. On the other hand, ignition of combustion or of a nuclear chain reaction are examples of uncontrolled amplification.)

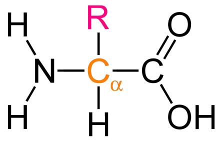
Abstractly, one minimizes some combination of the total binding energy between protein molecules and the ”information/selection cost” of DNA encoding for these molecules, but a mathematical rendition of Crick-Watson idea is still pending – no one of ”isoperimetric animals” cultivated by geometers for the last 3 000 years resembles icosahedral viruses.

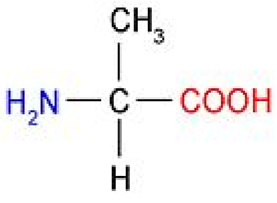
In the end of the day, the symmetry of viruses depends on the structural constrains imposed by the geometry of the *physical space*, which allows the existence of such improbable objects as icosahedra.

(The discovery of icosahedra in ≈ -400 by Theaetetus, who had hardly ever heard of viruses, can be only attributed to the unreasonable mathematical power of the brain visual processing system. Yet, brainless viruses had discovered icosahedra a couple of billions years before Theaetetus.)

5 Polypeptides and Proteins: Sequences, Folds and Functions.

A *polypeptide* is a polymer chain $A_1 \vDash_p A_2 \vDash_p A_3 \vDash_p ...$ made out of small basic unites - *amino acid residues.* There are 20 *standard amino acids*; most (not all) proteins in cells are composed exclusively of these 20.




alanin

A typical length of a chain is 100-300 residues (yet, reaching > 34000 in *titin* or *connectin*, $C_{132983}H_{211861}N_{36149}O_{40883}S_{69}$ – adhesion template for the assembly of contractile machinery in skeletal muscle cells.)

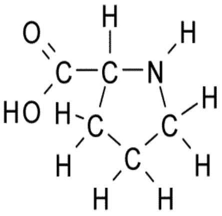
A "residue" is what remains of an amino acid after polymerization: the relatively strong covalent *peptide bond* $\vDash_p$ is formed between Carbon atom in each (but the last) amino acid molecule in the chain with Nitrogen atom in the next amino acid with a production of a water molecule.

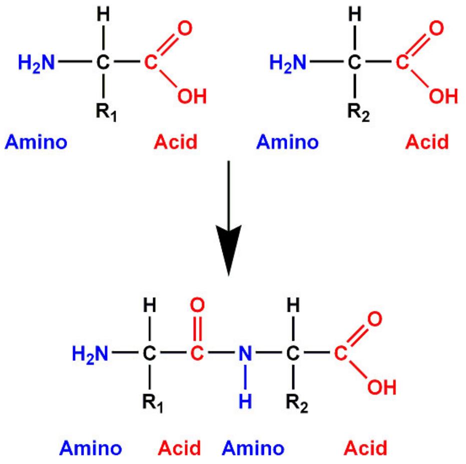
Immediately upon polymerization, a polypeptide chain, synthesized in the cell by the ribosomes and their "helpers" (making, conceivably, the most complicated chemical system in the astronomically observable universe) "folds" into a specific rather compact shape, called *protein*, held by additional *weak* binding forces between residues, mainly by hydrophobicity "pressure" and hydrogen bonds. (Seven out of 20 amino acids, including tryptophan and cysteine, are rather hydrophobic. They tend to conglomerate in the *protein hydrophobic core* with small exposure to the surrounding water.)

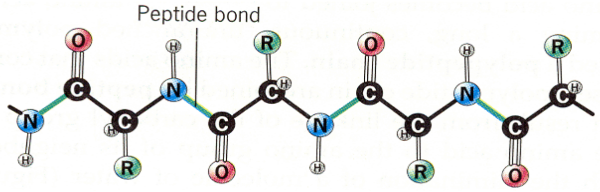
Some proteins are made of several polypeptide chains. For example, *haemoglobin* is composed of four $\approx$ 150-long subunit chains.

The beauty of proteins comes as much from a multitude of "little structures" within and around particular species of them as from yet unknown but vaguely felt general mathematical principles which underly their existence and properties. (Tens of thousand papers are dedicated to these "little ones" in the scientific literature, haemoglobin alone boasts $> 10^6$ entries in Google.)

Amino acids are small molecules, about 5Å in diameter (1Å= 0.1nm =







$10^{-10}m$), where 18 out of 20 standard amino acids are composed of Carbon, Nitrogen, Oxygen, Hydrogen and the remaining two (*cysteine* and *methionine*) also contain Sulfur. The smallest amino acid, *Glycine*, has 10 atoms in it and the largest – *Tryptophane* has 27 atoms.

Each amino acid contains the *main chain* of 9 atoms H_2N–CH-C-O_2H and a *side chain* (R-group) covalently bound to the central carbon atom called C_α in the main chain. As an exception, *Proline* (sketched above in black and white) has its 5-cyclic side chain also bound to N. Fifteen amino acid have tree-like side chains (with the single H for R in Glycine and CH_3 for R in Alanine) four have single cycles (with covalent bonds for the edges) in them and Tryptophan has two cycles.

Formally, a protein, or rather a polypeptide at this stage, is represented by a long word written on the backbone (which is a linear graph or a string with the peptide bonds for edges) in the letters of labeled graphs – the side chains of amino acid *residues.*

The linear structure of proteins was suggested in 1902 by Franz Hofmeister and a few hours later, at the same meeting, by Emil Fischer but many doubted that peptide bonds were strong enough to hold such long molecules together in the presence of thermal agitations. (On the surface of things, titin, for example, must have half life a few days. Is it more stable in conformation? Is it being constantly recycled? Does it function need stability? I have not looked enough into the literature.)

The beginning of ”era of sequencing” is landmarked with determination of the primary/sequential structure of the two polypeptide chains of *insulin* by Frederick Sanger in 1955 followed by Sanger’s method for sequencing RNA and

DNA in 1970's (This, in mathematics, can be compared to the turn in algebraic geometry and topology marked by the work by Jean Pierre Serre in 1950's.)

(Insulin is a very small *messenger protein* produced in pancreas of animals, secreted to blood and controlling intake of glucose by cells by binding to receptors on cell membrane. It consists of two polypeptide chains of 21 and 30 residues held together by week forces further stabilized by three covalent S-S bonds between Sulfur atoms in the side chains of Cysteine residues. There are two Cys in the 30-chain and four in the 21-chain; one S-S bond joins two residues in the latter and two bonds bridge the two chains together. Insulin is produced in two stages; first a cell synthesizes an insulin precursor, a polypeptide chain of $84 = 21 + 30 + 33$ residues and later the 33-segment is excised by a specific protein cleaving enzyme.)

The backbone chain of atoms in the polypeptide chain (with all but some terminal O and H omitted) looks as follows.

$$\mathrm{H_2N{-}C^{\alpha}{-}C'}\vDash_p\mathrm{N{-}C^{\alpha}{-}C'}\vDash_p\mathrm{N{-}C^{\alpha}{-}C'}\vDash_p \ldots \vDash_p\mathrm{N{-}C^{\alpha}{-}C'}\vDash_p\mathrm{N{-}C^{\alpha}{-}CO_2H}$$

The peptide bonds $\vDash_p$ are rather rigid with planar angles approximately 120°. The spacial flexibility of a polypeptide is mainly due to rotational freedom around N–C^α and C^α–C′ bonds; thus, the spacial *conformation* of an n-residue chain can be approximately parameterized by (a domain in) the Cartesian product of $2n$ circles with extra degrees of freedom coming from rotations around some bonds in the side chains, where the side chains also contribute extra degrees of freedom as well as "packing constrains" in space.

Most *native* (i.e. coming from living organisms) polypeptides assume, under specified conditions, a *unique* (?) spacial conformation, where some atoms, that may be far in the chain, come close together due to *weak interactions* (bonds) between them. Besides non-covalent weak bonds, stability of some proteins, e.g. of insulin, *ribonucleases* and of many snake poisons, is reinforced by covalent S-S bridges between sulfur atoms in cysteins.

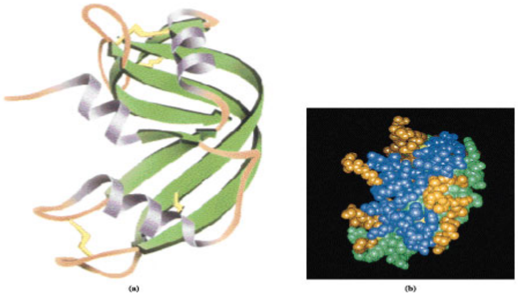
Bovine pancreatic ribonuclease-A (atomically $C_{575}H_{907}N_{171}O_{192}S_{12}$ which is represented schematically and atomically in the above picture) is a common enzyme in a biologist's lab. It uses 19 out of 20 amino acids, with eight cysteine involved in four S-S bridges and four *methionine* residues containing Sulfur atoms; only tryptophan is missing. It has hree α-helices and three β-hairpins where two of them form a four-stranded antiparallel β-sheet (defined below).

It is secreted by cows pancreas; its RNA cleaving activity is stable under up to 100°C, mainly due to the reinforcement of the enzyme structure by the S-S bonds. It is involved in digestion of RNA produced by micro-organisms residing in bovine stomach, where cellulose is broken down by symbiotic bacteria and protozoa.

Its amino acid sequence of 124 residues was determined in 1960 and the 3-D structure was solved in 1967 by the X-ray diffraction analysis.

It is commonly accepted by biologists that

all information required to specify the correct three-dimensional conformation of a protein is contained in its primary amino acid sequence. (The mathematical interpretations of these "all", "information", "required", "correct", etc. are by



no means unique.)

In bacterial cells, proteins should fold within at most a few minutes, since the life cycle of many bacteria is about 20 minutes.

Many (most?) proteins of length up to 150 – 250, when they are artificially unfolded by heating or by disturbing weak interactions with some chemical agent, spontaneously fold back (sometimes as fast as in a few milliseconds) to the native state when the conditions return to normal.

(Some people maintain that folding is an essentially co-translational process: a polypeptide chain, according to this point of view, would usually have hard time folding if it starts from a random position in the configuration space and/or will be not in contact with ribosomes and/or other protein complexes accompanying the translation process.)

This was established by Anfinsen's team in 1961 for Bovine Ribonuclease-A, where the reappearance of BRA's ability to degrade RNA was used as a witness of proper folding.

The uniqueness of foldings agrees with the existence of the *crystal* forms for many proteins since a *heterogeneous* mixture of nano-particles is unlikely(?) to make a crystal. Yet, there is a controversy about the universality, uniqueness and mechanisms of folding.

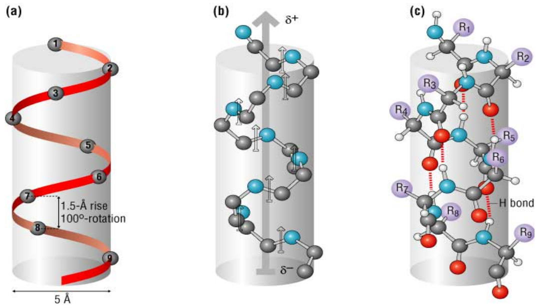
(Ambiguities and inconsistencies in presentation of ”basic facts” in biology are frustrating to a mathematician. For example, you can find statements in literature that insulin *can not renaturate* – to refold if unfolded; but some authors claim that the 33- chain is not crucial for bringing two other chains together and that under proper renaturation conditions the native insulin is obtained from a scrambled one with 25% yield that increases to 75% if the two chains are covalently linked.

But frustration turns into joy if you think of many interpretations of ambiguous statements allowing a variety of mathematical developments.)

How Proteins Fold. The basic folding patterns of proteins are called the *secondary structures* which are divided in two groups: *α-helices and β-sheets*; both are associated with the $\mathbb{Z}$-symmetry of the proteins backbones.

Helical structure was conjectured by William Astbury in the early 1930s on the basis of changes in the X-ray fiber diffraction of moist wool or hair fibers upon stretching; a detailed atomic model was worked out by Pauling, Corey and Branson in early 1950s.

(An essential component in the wool and hair, as well as in the outermost layer of cells of human skin, in the fingernails and in birds feathers, is *keratin.* Keratins make a group of fibrous proteins with helical molecules twisting around



each other with many S-S bridges between cysteine residues resulting in rigid structures. Human hair is about 15% cysteine. The characteristic smell of burning hair is due to the high presence of Sulfur in there.)

A typical helix contains about ten amino acids (about three turns) but some may have over forty residues. Helices are represented by rigid rods in schematic pictures of proteins.

Although the helix is formed by hydrogen bonding between the backbone residues, different amino acid sequences have different propensities for forming α-helical structure.

(For example, proline does not fit into helices because it misses an H-atom at N needed for the H-bond (with an O) and because its 5-cyclic side chain is bulky and rigid. At the other extreme, glycine, the smallest amino acid, also disrupts helices.)

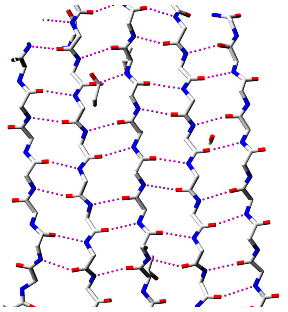
The β-sheet – the second form of regular secondary structure in proteins displays $\mathbb{Z}\oplus\mathbb{Z}$-symmetry. Such a sheet consists of several parallel or antiparallel (with respect to the $H_2N\to CO_2H$ direction of the backbone chain) *β-strands*, typically 5-10 amino acids long, connected laterally by hydrogen bonds and forming a pleated (often twisted) sheet. An example of (antiparallel) β is a *hairpin turn* (pictured in black and white below) where the two β-strand (almost) follow each other on the backbone.

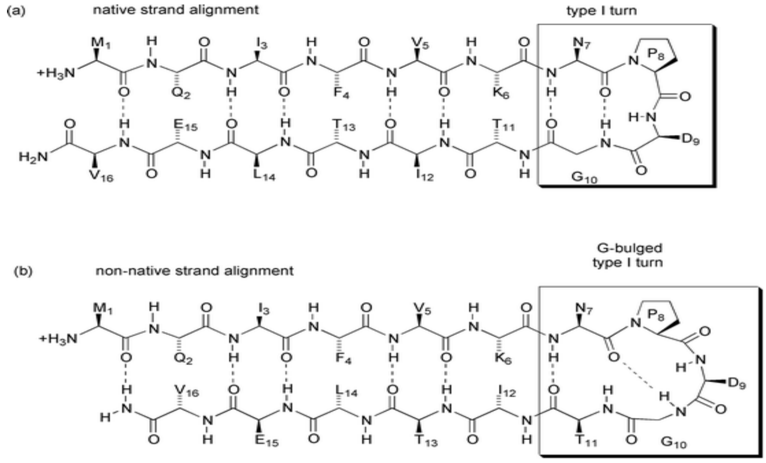
(The β-linkages may form between strands of different polypeptide chains making insoluble aggregates which cause, e.g., Alzheimer and mad cow diseases.)

The combinatorial structure defined by β-sheets is much richer, than what comes from α-helices: the arrangement of strands into sheets is a kind of *transformation* of the $\mathbb{Z}$-structure of the backbone to $\mathbb{Z}\oplus\mathbb{Z}$ of the sheets. The combinatorics of this transformation is determined by a subset $\mathcal{S}$ of particular segments S in the backbone, which represent the strands, and a graph on the vertex set $\mathcal{S}$ where the edges join the neighbor strands S in the sheets and where, moreover, the edges are labeled by the types of these neighbor relations, e.g. being parallel or antiparallel.

The combined combinatorics of this graph (which is by itself not very informative) with the order structure on the strands in the backbone displays, in many instances [4], tree-like (nested) patterns similar to those in *parsing* of sentences into words and phrases in *context free languages*.

Besides pure combinatorics, the rigid "rods" (for α-helices) and "plates" (for β-sheets), joined by rather flexible *loops*, make a particular arrangement (of





the backbone of a protein) in the 3-space, called the *super-secondary/tertiary structure* of a protein.

One still lacks a comprehensive formal language for describing such structures as was emphatically pointed out to me by Arthur Lesk on several occasions (See [20] for a new mathematical approach.)

Protein Binding. Most of functions of proteins in cells depends on *specific binding* of a protein P to another molecule M or to a particular class of molecules.

For instance, several protein molecules make a viral capsid (shell) by binding one to another.

Signaling/messenger proteins, e.g. insulin, bind to specific receptor proteins on the exposed membranes of cells.

Regulatory proteins bind to specific segments of DNA, thus enhancing or suppressing the transcription of RNA.

(Proteins are rather sticky due to non-uniformity of electric charges on their exposed parts and tend to non-specifically bind one to another. Amazingly, this does not mess up the normal activity of the cell.)

Catalysis. Certain proteins, called *enzymes*, speed up chemical reactions. (About 4,000 such reactions are known in cells). For example, let M be a molecule where some specific bond is metastable: the energy is released when this bond is broken, such as the peptide bond in the water environment for instance or the phosphodiester bond in RNA between nucleotides.

What prevents this break to happen spontaneously, or rather makes it highly improbable, is a potential barrier – a "mountain range" in the configuration space Y of M around a metastable state $y_0 \in Y$ which prevent the "flow of energy" from y_0 – a local minimum of the energy function on Y – to the true minimum y_{min} of the energy.

A catalyst provides a channel across this barrier (or widens the existent but a very narrow and/or long twisted channel) which allows a passage from y_0 to y_{min}. This may be implemented by different chemical mechanisms (which are still not fully understood [16]).

For example *ribonucleases* bind to RNA molecules at specific places and "cleaves" the covalent bonds between the nucleotide residues.

Enzymes of our digestive system, such as *pepsin* released in the stomach and *trypsin* produced in the pancreas break peptide bonds in proteins at certain amino acid residues which are specific for particular enzymes.

Catalase, which is contained in tissues of organisms exposed to oxygen, e.g. in your saliva, splits H_2O_2 into water and oxygen. This is an instance of a *kinetically perfect enzyme* limited only by the diffusion rate of the substrate. (If you spit into a water solution of H_2O_2, the liquid starts bubbling with oxygen.)

What are the configuration space representations depicting the catalytic functions of these enzymes?

Enzymes are extremely good at what they do: some speed up reactions 10^{17}-10^{18} fold. (This is said about *orotidine 5'-phosphate decarboxylase* [15] but I failed to find a reference to how one measures the rate of non-catalized reactions in such cases.)

6 Energy Landscapes and Protein Problems.

Let $X = X(P)$ be the configuration space of a given polypeptide chain P in $\mathbb{R}^3$ and let $E : X \to \mathbb{R}$ be the total energy "summing up" the weak interactions between the (atoms in the) residues.

Can one find the "ground state" $x_{min} = x_{fold}$ which minimizes the energy and corresponds to the folded protein? (After all if the nature does it in a few seconds, why a mathematician can not do it?)

The folding process in this $X(P)$-model corresponds to the down stream gradient flow of E where the orbit of every $x \in X$ eventually arrives at x_{min}. But E may have lots of local minima: their number is likely to grow *exponentially* with the dimension of X that is (roughly) proportional to the number N of residues in P. If so, a protein can not fold reasonably fast, if at all, for most E.

The problem of "shallow" local minima disappears in a more realistic model where thermal fluctuations are incorporated into the picture. Now, the folding process is represented by a *random walk* which is biased according to E, where the probability density of a point $x_1 \in X$ moving in time δt to a nearby point x_2 is proportional to $(\delta t) exp(E(x_1) - E(x_2))$. (Roughly, this corresponds to smoothing the function E; thus, erasing insignificant local minima.)

The "ground/folded state" is represented in the randomized flow by a "small" neighbourhhod U_{fold} of $x_{min} = x_{fold}$ such that our random walk stays in U_{fold} "most" of the time. (A comprehensive model must incorporate statistical dynamics of water molecules which is crucial for the "hydrophobic component" of the "folding force".)

Yet, one may expect that a "typical E" would have many local minima with rather wide/deep attraction basisins, so that x would keep jumping from one U to anotrher. In other words, the sublevel $U(\varepsilon) = E^{-1}(-\infty, E(\varepsilon)) \subset X$ with the stationary probability masure $1-\varepsilon$ may have many connected componets spread over X and not localized at all around any x_{fold}. And even if $U(\varepsilon)$ is localized, the time of arrival to it from a "random" $x \in X$ may be too long. There must be something special about E which allows (fast) folding.

It is hard to say more as we do not have sufficient understanding of the connectivity properties (not to speak of higher dimensional homology) of sublevels of "random" functions on high dimensional spaces. Something in this regard is provided by the *percolation theory* but this concerns *limits* of spaces of *fixed* dimension where the size of a domain goes to infinity (e.g. in lattices) or, conversely, where only the dimension goes to infinity (as for the n-cliques and n-cubes.)

Here, however, both the size of the space $X = X(P)$ as well as its dimension grow with the length N of P, and where something interesting happens not so much in the limit but at specific (large but not very large) values of parameters. (Too long polypeptide chains, well above 500 residues, do not fold, if at all, into anything like compact protein globules, but some large proteins may consist of several independently folded globular domains while some, such as titin and keratin, have a fibrous structure.)

Apparently, the the backbone degrees of freedom contribute to the size of X as well as to its dimension, while the side chains' degrees of freedom contribute only to $dimX$.

If the size of X is small on the "oscillation scale" of E (which is controlled, to a large extend, by the average value of the gradient of E) and if the dimension of X is large, one expects a unique large cluster (connected component) of a (low energy) sublevel which covers most of (measure of) X.

But if X is large compared to $dimX$ then one expects many local minima and high disconnectedness of sublevels (of comparable energy) of random functions E.

In the latter case one can not expect folding of "generic" polypeptides, but one can imagine that some special polypeptides fold similarly to how crystals and micelles are formed – the folding process is directed by something like a "riverbed with tributaries" in $X = X(P)$ which channels the gradient flow toward X_{fold}, where specific patterns in (combinatorics of "tributaries" of) this riverbed, say $R = R(P)$ correspond to interactions between particular "important" (groups of) residues in P and where these patterns were selected by Nature for *native* proteins P in the course of evolution. (In the case of protein crystals the role of selector is taken by a crystallographer.)

Can one make this mathematically precise?

Can one get an insight into R for specific (families of) proteins on the basis of their conformations and/or of their biochemical properties?

Can one artificially design proteins which would fold by controlling their R?

A more realistic mathematical problem is that of finding a class of models of high dimensional stochastic gradient-like systems which may be far from real proteins but where the above questions have positive answers.

But should one stick to this huge space $X(P)$ anyway? After all, one has no experimental access to all of X (even the "theoretical existence" of such "full X" is debatable) but rather to some quotients of X corresponding (to sets of) particular observables (functions) on this space, where the random dynamics on X defined by the probabilities of transitions from one domain in X to another within an infinitesimal time interval δt naturally induces such dynamics on every quotient space Y of X. (This suggest that some projective limit of the observable quotients of X may replace X.)

In particular, the connectivity of sublevels of a function E on a topological space X is encoded by such a quotient – the *sublevel tree* $T = T(X, E)$. This is a tree with a continouos map $\tau : X \to T$ and a function $E_T : T \to \mathbb{R}$, such that $E = E_t \circ \tau$ and where the map induced by τ from the set of connected components of every t-sublevel of E (i.e of the subset $E^{-1}(-\infty, t] \subset X$) onto the t-sublevel set for E_T is one-to-one.

The tree T comes with a natural *metric* on it for which $E_T : T \to \mathbb{R}$ is *isometric on every edge* of T. Besides it carries the stationary measure of the random walk induced on it from X.

(Actually, every measure on X induces a measure on T. It is tempting to use the product of the angular measure of rotations around simple covalent bonds in the residues in P but it is unclear to me if this make much sense.)

How much of protein properties, and of molecular assemblies in general, can be expressed in the language of these trees T?

There is a natural *convolution product* operation on such trees (which is again a tree), say $T_1 \star T_2$, which corresponds to non-interracting systems, where $T_1 \star T_2$, equals the sublevel tree for the sum of energies, $E_1 + E_2$ on $T_1 \times T_2$. (This generalizes/refines the convolution of measures on $\mathbb{R}$, where a somewhat different kind of object – a general *graph* instead of tree – comes up if one uses *levels* instead of sublevels of E.)

If the systems do interact, then the full energy on the product of the respective spaces is written as $E_1 + E_2 + E_{1,2}$, where the interaction term $E_{1,2}$ is, usually localized on a certain relatively small part of the product space.

For example, the effect of a catalyst, e.g. an enzyme, can be, apparently, seen in terms of this $E_{1,2}$ on the products of full configuration spaces but the T-quotients may be too small for this.

What are the smallest quotients of the full configuration spaces which would allow an adequate geometric description of enzymatic catalysis? How much does this depend on the type of an enzyme P?

The above mathematical problems are compounded by physical ones [6].

1. The weak interaction energies between (atoms in) residues are known only approximately: their quantum mechanical derivation is far beyond our computational capability and nor direct experiment can detrmine interatomic/molecular interaction with a sufficient precision.

2. The total interaction energy E of a polypeptide is *not* the sum of the pairwise residue interaction energies.

However, even if non-binary and non-strictly additive, the interaction energy is a relatively simple (unknown) function of (the sequantial composition of) a protein which, probably, can be encoded with a reasonable accuracy by something like $10^4 - 10^6$ bits of information. (For example, if the energy were the sum over the pairwise residue interaction energies, $E = \sum E_{ij}$ which we wanted to estimate up to an ε, we would need about $20^2 \log(N/\varepsilon)$ bits for all proteins with N amino acid residues.) On the other hand one has much high throughput experimental data on (properties of) proteins, where each experiment carries at least one bit.

What is a general mathematical "parameter fitting" method(s), which, when applied to proteins, could provide (an effective version of) the total inter-residue interaction energies?

(Such approach is pursued in bioinformatic but it does not seem to incorporate the biochemical data available, e.g. on the calorimetry of proteins (un)foldings and/or on protein-protein$'$ binding, say, for protein-*immunoglobulin.*)

What Percentige of Polypetride Chains fold? The number of, say, N-long amino acid sequences is $20^N \approx 10^{1.3N}$ but the overwhelming majority of polypeptide chains, probably, do not make anything even vaguely reminding proteins. (Random sequences of letters make no sentences in the English language either.)

But is there any way to describe the "subset" (space) $\mathcal{P}$ of all "*conceivable* proteins" P in the full sequence space ? What is the cardinality of this "space" $\mathcal{P}$?

Write the cardinality of the set of "conceivable" protein sequences of length N as $20^{\sigma N}$ for $0 < \sigma < 1$ and think of $1 - \sigma$ as "codimension" (coentropy) of the space $\mathcal{P}$, where the "dimension" of the space of all sequences is normalized to 1. In other words, $1 - \sigma$ represents the "number of equations" or constrains which a sequence has to satisfy in order for P to behave "protein-like", where the propensity to fold is an essential ingredient of being "protein-like"

In reality, σ depends on N as well as on a particular protein P – a point in $\mathcal{P}$ where σ is evaluated.

The full space $\mathcal{P}$ is too large to be studied experimentally, but one can evaluate the proportion of proteins P' obtained from a given P by a few mutations, e.g. by substitutions of some residue by another one in the sequence, such that P' still folds.

A lower bound on such "local" $\sigma(P)$ for native proteins can be extracted from the data on the mutation rate of proteins estimated by comparing sequences of homologous proteins of different organismas [19].

Namely, let $r = r(P)$ be the mutation rate of P and $R = R(P)$ be the rate of "fictitious mutations", usually called *synonymous* mutations of DNA which do not change the corresponding amino acid (because of the redundancy of the genetic code). Then he ratio r/R provides a plausible lower bound on σ, since the mutations must not only preserve the folding but also the functionality of P.

The evolutionary data suggest that the ("folding component" of the) value of σ is somewhere around $1/2$ (rather than being close to 0 or to 1; overlaps of some viral genes suggest $\sigma > 2/3$ for small proteins) but it remains unclear how much a particular value of r/R, which greatly varies across different families of proteins, depends on the type of folding (e.g. on the super-secondary structure of a protein P) and how much is due to the functional constrains. (Possibly, one may extract some information by comparing the mutation rates of proteins with similar structure versus such rates for proteins with similar functions.)

For example, a protein which specifically interacts with several neighbor proteins, is expected to have small ratio r/R but this does not necessarily effects the "folding componenet" of its σ. This ratio is sometimes effectively zero, e.g. for the *histone protein* H4 in chromosomes (which is coded by several genes), and then the meaning of r/R becomes questionable. (This histone's conservatism is usually explained by the close structural association of its molecules with neighboring protein molecules and DNA in *chromatin*, but this reasoning does not apply to highly mutable viral capsid proteins.)

Interestingly enough, already in 1904, George Henry Falkiner Nuttall observed that rabbits' antibodies/immunoglobulins developed against human blood proteins were equally well precipitating the blood serum of African apes (but no so of Asian apes), thus showing close similarity of the corresponding proteins prior to even any idea of their sequences. Then the comparative immunology was used by Morris Goodman in 1961 for establishing evolutionary relationships among primates.

However, there is no general (semi-mathematical) approach combining biochemical and sequential/evolutionary data for evaluation of essential structural properties of proteins such as the relative roles of structural and functional constrains reflected in r/R. (But the evolutionary comparison is systematically used in bioinformatics for predicting protein conformations by their sequences.)

One may expect that the main contribution to σ can be expressed by pairwise correlations between residues at specific positions on the chain which somehow influence one another in the conformation but proving (and even stating) this mathematically seems hard; one is tempted to look at similar more approachable models of "design" and/or of "evolution" of "stochastic gradient flows" of functions on high dimensional spaces. (Probably, there are tractable model problems in the percolation theory.)

Everything we presented in this paper hardly scratches the surface of what is known about crystals, cell membranes, virus capsids and proteins, where understanding structure and molecular function of proteins constitutes the first step in the solution of the main biological problem of the "sequencing era".

Describe the arrow *genotype* $\rightsquigarrow$ *phenotype*, where "genotype" is given by the genome of an organism, possibly "ornamented" by some epigenetic data (e.g. *methylation* of some bases and positions of some regulatory proteins on DNA).

This problem can be divided into several parts.

1. Determine the domain of definition of the map "$\rightsquigarrow$" by finding a realistic (mathematical) description of the subset $\mathcal{G}$ of sequences which may serve as viable genomes G, of "conceivable organisms", where this description must be expressed in the language of sequences.

There may be several such description on different level of precision, where such a description is supposed to be only approximate with a balance between the degree of approximation/precision and "mathematical simplicity/complexity" of the description.

We have briefly addressed this for individual proteins P, where a protein sequence and function is determined (modulo alternative splicing, translational regulation and posttranslational modifications) by the DNA code for P.

2. Formally describe "phenotype", let it be only approximately, on several levels of precision.

This is non-trivial even for an ndividual protein P where its "phenotype" includes both the structure (conformation) of P as well as its function(s) and where the simplest to describe (but often hard to determine) among functions are the protein's binding and enzymatic properties.

3. Represent (possibly, only approximately/statistically) the "space of phenotypes" or, rather, a significant (sub)quotient of this space, as a quotient of the space of (possibly, slightly annotated) genomes by some equivalence relation with an effective description of this relation in the sequential language. (This is similar to describing the "real world" as a quotient space of the "space of sentences" of a natural language.)

In particular, define and evaluate some numerical measure(s) of "redundancy of the map $\rightsquigarrow$" associated to the cardinalities of the fibers of this map.

A special and a more realistic subproblem is doing this in a "neighborhood" of an individual native protein P, where this redundancy may be expressed by an equivalence relation on the set of amino acid sequences close to those of P and giving a protein P' similar to P. In particular, we want to say "something interesting" on the domain of continuity of the map $\rightsquigarrow$ and on jumps of $\rightsquigarrow$ at the discontinuity points.

A mathematician's role in solving these problems may consist in designing a "parameter fitting scheme" for determination of a "mathematically/logically simple(st)" arrow $\rightsquigarrow$ (or fragments of $\rightsquigarrow$) compatible with (constrained by) two kinds of data.

A. The data (e.g. obtainable from protein data banks) on sequences, structures and functions of proteins.

B. Known physics/chemistry of proteins. (These data need a preliminary uniform formal representation.)

Besides the above, there are two other general problems in molecular biology with a mathematical tint to them: *combinatorial design of high throughput experiments and description of the "moduli space(s)" of proteins (and genomes).*

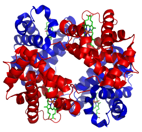
The latter is associated to the map from the "protein tree of life" into the protein space $\mathcal{P}$, where the slow dynamics of evolution shapes the fast dynamics of folding and of enzymatic catalysis. (The "tree of life" is not a "bare tree": the horizontal gene transfer, artificial construction of chimeric proteins, the position and the number of genes on DNA which code a protein, protein interactions, etc., add much extra structure to this "tree".)

Understandably/regretfully, our present view on "protein problems" falls into a traditional mathematical framework. Hopefully, an attempt to solve them may lead to something new and unexpected.

The *Bar-headed Goose* flies up to 10 km when migrating over the Himalayas. Its haemoglobin allows the goose to breath −50°C air at 25% sea level density. This haemoglobin differs from the haemoglobin of its lowland relative by four amino acids, where, arguably, only one of the substitutions, Proline $\mapsto$ Alanine, contributes to the jump in the goose haemoglobin oxygen affinity and in the goose' ability to fly high.

The Bar-headed Goose has two kinds of Haemoglobin in its blood, where only one of them has the elevated Oxygen affinity; the presence of the second one allows goose' adaptation to low altitudes.

Even more remarkably, *Rüppell's vulture*, (G. rueppellii) has four different types of haemoglobin in the blood [13].



(On November 29, 1973 an aircraft collided with a bird over Côte d'Ivoire at altitude 11,300 m as recorded by the pilot shortly after the impact. The plane landed safely at Abijan. Sufficient details in the remnants of the bird allowed their identification as G. rueppellii [28].)

7 Bibliography.

References

[1] Andersson S, Larsson K. Larsson M. Virus Symmetry and Dynamics, $http://www.sandforsk.se/sandforsk-articles/all.all.articles.htm$

[2] Alan J. Cann, Principles of Molecular Virology, Science, 2005

[3] Caspar, D. L. D. and Klug, A. (1962) "Physical Principles in the Construction of Regular Viruses" Cold Spring Harbor Symposia on Quantitative Biology XXVII, Cold Spring Harbor Laboratory, New York. pp. 1-24.

[4] Chiang Y.S., Gelfand T.I., Kister A.E., Gelfand I.M. New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. Proteins, 2007.

[5] Crane, H. R. Principles and problems of biological growth. The Scientific Monthly, Volume 70, Issue 6, pp. 376-389. (1950) (This article is often referred to but I could not see it since it is not freely available on the web.)

[6] Finkelstein, A. V. : Protein Physics : A Course Of Lectures. Academic Press, 2002

[7] M. Gromov, Isoperimetry of waists and concentration of maps, Geom. Funct. Anal. 13 (2003), 178–215 GAFA, Vol. 13 (2003) 178 215.

[8] Gromov, M., Mendelian Dynamics and Sturtevant's Paradigm. In Geometric and Probabilistic Structures in Contemporary Mathematics Series: Dynamics, (Keith Burns,, Dmitry Dolgopyat, , and Yakov Pesin editors), American Mathematical Society, Providence RI (2007)

[9] Larry Guth, Minimax problems related to cup powers and Steenrod squares, GAFA 2009, to appear.

[10] Mike Howard, Introduction to Crystallography and Mineral Crystal Systems.

http : //www.rockhounds.com/rockshop/xtal/index.html

[11] W. Helfrich, Z. Naturforsch, Bending energy of vesicle membranes: General expressions for the first, second, and third variation of the shape energy and applications to spheres and cylinders, Phys. Rev. A 39, 5280 - 5288 (1989).

[12] Kushner, D.J. Self-Assembly of Biological Structures, Bacteriol Rev. 33:302-345. 1969.

[13] Wen-Hsiung Li, Molecular Evolution, Sinauer Associates, Sunderland, MA, USA. 1997.

[14] Klaus Mainzer, Symmetries of Nature , Walter De Gruyter, NY, 1996,

[15] Miller BG, Wolfenden R. (2002). "Catalytic proficiency: the unusual case of OMP decarboxylase.". Annu Rev Biochem. 71: 847885.

[16] Kenneth E. NeetDagger, Enzyme Catalytic Power Minireview Series, J Biol Chem, Vol. 273, Issue 40, 25527-25528, October 2, 1998.

[17] Christo N. Nanev, How do crystal lattice contacts reveal protein crystallization mechanism? Cryst. Res. Technol. 43, No. 9, 914 920 (2008)

[18] Masayoshi Nakasako, Water-protein interactions from high-resolution protein crystallography. Philos Trans R Soc Lond B Biol Sci. 2004 August 29; 359(1448): 11911206.

[19] Laszlo Patthy, Protein Evolution, Blackwell Science, 1999.

[20] Penner, R. C.; Knudsen, Michael; Wiuf, Carsten; Ellegaard Andersen, Joergen, Fatgraph Models of Proteins, eprint arXiv:0902.1025.

[21] Bernard Testa, Joachim M. Mayer, Hydrolysis in Drug and Prodrug Metabolism: Chemistry, Biochemistry, and Enzymology, Science, 2003.

[22] Radzicka A., Wolfenden R., RATES OF UNCATALYZED PEPTIDE BOND HYDROLYSIS IN NEUTRAL SOLUTION AND THE TRANSITION STATE AFFINITIES OF PROTEASES Journal of the American Chemical Society 1996Vol.118(No.26)

[23] Vliegenthart G. A., Gompper G, Mechanical Deformation of Spherical Viruses with Icosahedral Symmetry, Biophys J. 2006 August 1; 91(3): 834841.

[24] Reddi K.K. TOBACCO MOSAIC VIRUS WITH EMPHASIS ON THE EVENTS WITHIN THE HOST CELL FOLLOWING INFECTION in Advances in virus research, Volume 17 Kenneth Manley Smith ed. 1972 - Science.

[25] *http : //pathmicro.med.sc.edu/mhunt/intro − vir.htm*,
http : //www.microbiologybytes.com/introduction/structure.html,
http : //web.uct.ac.za/depts/mmi/stannard/virarch.html)

[26] Google: timeline history icosahedral virus.

[27] Conformational Proteomics of Macromolecular Architecture: Approaching the Structure of Large Molecular Assemblies and Their Mechanisms of Action(With CD-Rom) (Paperback) by R. Holland Cheng (Author), Lena Hammar (Editor) 2004, a historical survey by Morgan.

[28] *http : //elibrary.unm.edu/sora/Wilson/v086n04/p0461 − p0462.pdf*