Mathematics of Life Spaces: Continuation of the 2018 Large Dimensions Course

Misha Gromov

September 25, 2019

1 Lecture 1

The grandfather of all spaces, denote it A^B, is the space of

all conceivable texts written on a background set/space B
in the letters from an alphabet A,

or, in mathematicians parlance,

the space of maps $B \to A$.

In physics, which is dominated by numbers, and in the branches of mathematics oriented toward physics, A is the set of real numbers and the corresponding multidimensional space is the Euclidean one denoted $\mathbb{R}^N$, where this N is also a number, now a natural rather than a real one, that stands for the cardinality (possibly infinite) of some background space B.

Then this N, the cardinality of a *non-specified* B, is identified with the ordinal number N, that is represented by the set $\{1, ..., N\}$; thus making elements of $\mathbb{R}^N$ written as *sequences of real numbers* $(a_1, ..., a_i, ..., a_N)$, which are called N-*vectors*.

The space $\mathbb{R}^N$ with the *Euclidean* (Pythagorean) metric, where the distance between sequences $(a_i), (b_i) \in \mathbb{R}^N$, $i = 1, ..., N$, is

$$dist\,((a_i), (b_i)) = \sqrt{\sum_i (a_i - b_i)^2},$$

is *implausibly symmetric*:

not only the the action of the isometry group on $\mathbb{R}^N$ is *transitive*, that is any point/vector x can be moved to another y by an isometry of the space $\mathbb{R}^N$, but this is also true for arbitrary k-tuples of points $X = \{x_j\}$ and $Y = \{y_j\}$:

whenever

$$dist(x_{j_1}, x_{j_2}) = dist(y_{j_1}, y_{j_2}) \text{ for all } j_1, j_2 = 1, ..., k,$$

there is an isometry of the ambient Euclidean space $\mathbb{R}^N$ which moves X to Y.

For instance, the configuration space of N points/particles in the 3-space, seen as $\mathbb{R}^{3N} = (\mathbb{R}^3)^N$, comes along with an enormous $\left(\frac{3N(3N-1)}{2} + 3N\right)$-dimensional isometric symmetry group[1] most of which is invisible to an eye blind to the Pythagorean theorem.

[1] The rotations of $\mathbb{R}^{3N}$ make what is called the *orthogonal group* $O(3N)$ of dimension $\frac{3N(3N-1)}{2}$ and there are also $3N$ parallel translations.

In physics, the same exponent **2**, which is responsible for this miraculous symmetry, plays a similar role, where it enters, for instance, the formula for the kinetic energy:

$$energy = \frac{mass \times velocity^{\mathbf{2}}}{2}.$$

Modulo the omnipresence of the magnificent **2** in mathematics and in physics alike, and a couple of similar miracles,

the effectiveness of mathematics in physics shouldn't strike you as "unreasonable".

But why, as some say,

is mathematics unreasonably ineffective in biology?

Here one has to pause and ask himself/herself:

What kind of mathematics? If it is the above kind of mathematics which is modulated by physics and saturated by numbers, why should it be adapted to an effective description of the tangle of live structures?

These structures are as much miraculous as the ones you see in the physical worlds but these are different kind of miracles. A traditional physicist's mind's eye is blind to them and the physical style math. is no match for them.

Imagine, your visual system were trained on a planet bare of life, where the only patterns to learn were clouds in the sky, waves in the water and random arrangements of rocks in the planes.

Then the mental image in your brain of a moving object of a kind of we see in Life, be it a running elephant or a running truck, would, probably, register as a solitary wave, if at all.

And human psychology apart, the traditional tools of "physical mathematics" are too refined, polished and smooth to catch the essence of live things. Doing this is would be like trying to drive a car in the world without friction: mathematics is beautiful, logic is perfect but the car has no intention to move, the steering wheel is too slippery to turn and, in a few moments, bolts and nuts get loose and the car dissociates into pieces.

(Probably, one can rigorously prove that no life-like structure would be possible in the world governed by ideally symmetric math, similarly to impossibility of assembling an electronic device from "perfect" materials – all satisfying *the Ohm law*.)

Although this kind of mathematics reduces to a bunch of numerical trivialities when it comes to such Life spaces as

the space of bacteria in human guts,
the space of sentences in a language,
the space of ideas in one's brain,

there are remarkable instances of mathematical spaces in Life, where "physical mathematics" does tell you something substantial, e.g.

the spaces of proteins molecules in solvents

and

phylogenetic trees of protein sequences

We say a couple of words about these below, and then return to such spaces in lectures ???.

1.1 Proteins and Proteins Folding

Proteins come to life[2] as long molecules build of 20 (sometimes 21 or even 22) units, called *amino acid residues*, where the number of these is, for (*globular*) proteins we speak about now, is in the range of 100-300 residues.

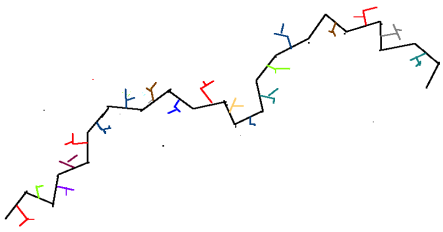


Figure 1: Schematic picture of a heteropolymer with homopolymeric main chain

Formally, this a string in 20 symbols, kind of a word or a sentence in 20 letters, where the chemistry of this is as depicted below.

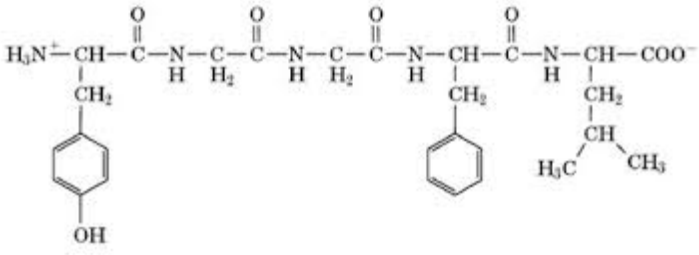


Figure 2: Chemical picture of an amino acid chain

Then, granted a proper temperature, acidity etc, these chains "fold" in fraction of a second into compact densely packet fairly stable potato shaped blobs of protein molecules.

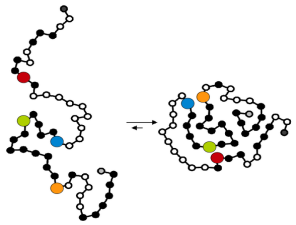


Figure 3: protein folding and unfolding

This folding process, that transforms

$$\text{sequence} \rightsquigarrow \text{structure}$$

[2]They are synthesized by *ribosomes* in the cells.

is a key elementary step toward Life:

informations encoded by a sequence in 20 letters turns into something corporeal: a structurally organised functional protein.

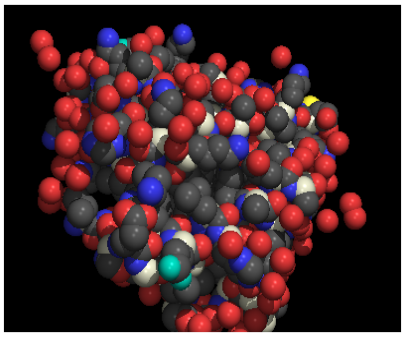


Figure 4: Spatial model of a folded protein molecule

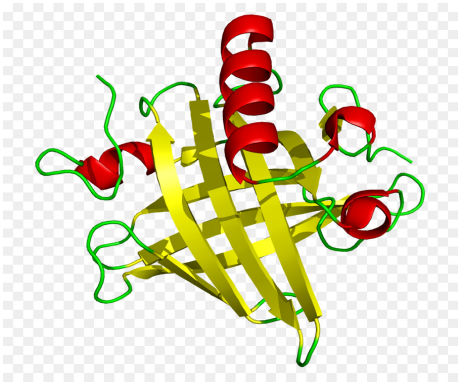
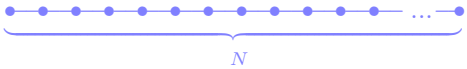


Figure 5: Schematic protein structure.

But given a sequence, one can't theoretically tell if the protein folds, if it does, than how fast, and what will be the shape of the resulting compact molecule.

1.2 Self Avoiding Random Walks

A (very) rough mathematical model of an *unfolded* protein molecule in a solvent is that of the *N-step self avoiding random walk*, that is a random imbedding of the standard linear graph



to the 3D lattice, where "random" refers to the *uniform measure* on the set of imbeddings, i.e. with equal weights assigned to all embeddings. [3]

The most amazing thing about this model is that, despite being a gross over-simplification of the structure of unfolded proteins, which is hardly considered as problem at all by the protein folding community, it remains mathematically inaccessible: none of the "intuitively obvious" properties of these "walks" have been rigorously proved so far.

It is *conjectured* that the diameter of a random N-chain in the plane grows with N as $\sim N^{\frac{3}{4}}$ and the expected growth of this diameter in the 3-space is something like $N^{0.588}$,

But it is unknown if the diameter of random N-chain in the 3-space grows at last as $\sqrt{N}$, i.e. no slower than what happens for the ordinary random walk.

What is obvious is that, since the ball of radius R in the lattice $\mathbb{Z}^d$ can't contain more than $const \cdot R^d$ distinct points, the diameters of all self avoiding N-chains in $\mathbb{Z}^d$ satisfy

$$diam_N \geq const_d \cdot N^{\frac{1}{d}}.$$

[3] As we shall explain later on, the very applicability of the concept "random", accompanied by the physicist's intuition attached to it, remains problematic for this model.

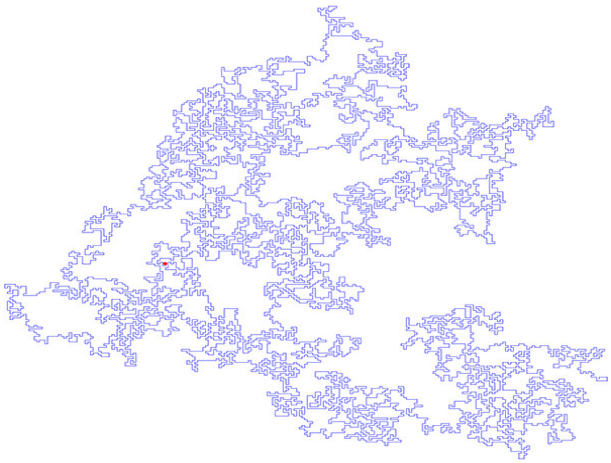


Figure 6: Self avoiding walk in $\mathbb{Z}^2$

But what is proven is that

the expectation of the squared displacement of the self avoiding walk in $\mathbb{Z}^d$ *satisfies*

$$\sqrt{\mathbb{E}||x_1 - x_N||^2} \geq const \cdot N^{\frac{2}{3d}}, \text{ [4]}$$

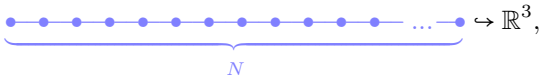
which, observe, even for $d = 2$*, doesn't imply the obvious bound* $\sqrt{\mathbb{E}diam_N^2} \geq const \cdot N^{\frac{1}{2}}$. [5]

And, as far as the lower bound on the diameter $diam_N$ is concerned, it is proven, for self avoiding walks in $\mathbb{R}^d$ for all $d \geq 2$, that this diameter grows sublinearly, in the sense that

the probability of $diam_N \geq \varepsilon N$ *is at most* $(1 - \delta(\varepsilon))^N$ *for all* $\varepsilon > 0$ *and some function* $\delta(\varepsilon) > 0$.[6]

,

What we want to understand, however, is not only and not so much the geometry of the images of random embeddings or "walks", depicted as



$\hookrightarrow \mathbb{R}^3$,

but rather of

the space $\mathcal{S}$ of all such embeddings

also called *self avoiding walks* in the present context.

In what follows, we return to the geometric setting closer to what happens to proteins, where we deal with self avoiding walks in all of $\mathbb{R}^3$ instead of the integer lattice $\mathbb{Z}^3$ and where all $\bullet$ in  stand for small rigid balls, rather than dimensionless pointsand where this $\mathcal{S}$ is

[4] Madras N. (2014) A lower bound for the end-to-end distance of self-avoiding walk. Canad. Math. Bull. 57, 113 -118.

[5] It may seem obvious that $\mathbb{E}diam_N^2$ can't be *significantly* greater than $E||x_1 - x_N||^2$, but, apparently, this also remains only conjectural.

[6] Hugo Duminil-Copin, Alan Hammond (2012), *Self-avoiding walk is sub-ballistic.* arXiv:1205.0401v1.

a fairly complicated subset in a larger but much simpler space $\mathcal{W}$ of all "walks" in the 3-space $\mathbb{R}^3$.

Formal Definitions of $\mathcal{W}$ and $\mathcal{S}$. A (unitary) N-step walk W in the 3-space is a sequence of points $x_1, x_2, ..., x_N \in \mathbb{R}^3$, such that

$$dist(x_i, x_{i+1}) = 1 \text{ for } i = 1, 2, ...N-1.$$

Thus the space of these "walks" is a submanifold

$$\mathcal{W}_N = \mathcal{W}_N(1) \subset \mathbb{R}^{3N}$$

of codimension $N-1$.

Equivalently, $\mathcal{W}_N$ can be described as the *product space,*

$$\mathcal{W}_N = \mathbb{R}^3 \times \underbrace{S^2 \times ... \times S^2}_{N-1},$$

where the factor $\mathbb{R}^3$ represents the positions of the first vector x_1 from the string $x_1, ..., x_n \in \mathbb{R}^3$ and where the unit vectors in the ith sphere $S^2 \subset \mathbb{R}^3$ in the product are the differences $x_{i+1} - x_i$, which, for $i = 2, ...N-1$, can be thought of as the "2-dimensional angles" between the segments $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$.

Then the the subspace of self avoiding walks

$$\mathcal{S}_N = \mathcal{S}_N(\varepsilon) \subset \mathcal{W}_N$$

is obtained by removing those W from $\mathcal{W}$, where the W-images of some balls/beads • from the chain intersect in the 3-space, where they are mapped by W.

More formally, the subspace $\mathcal{S}_N = \mathcal{S}_N(\varepsilon) \subset \mathcal{W}_N \subset \mathbb{R}^{3N}$ is distinguished by the inequalities

$$dist(x_i, x_j) \geq 2\varepsilon, \ i \neq j,$$

where $\varepsilon < \frac{1}{2}$ is the radius of • in the chain •—•—•—•—•—• ...[7]

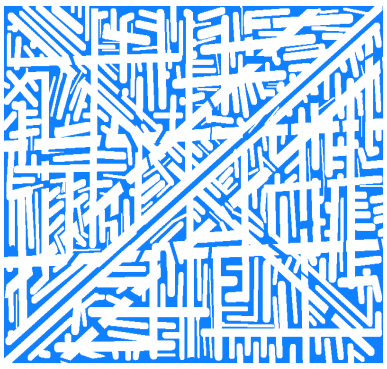


Figure 7: The square represents $\mathcal{W}$, the white strips correspond to the positions of pairs of intersecting • and the blue complementary to these strips depicts $\mathcal{S}$

Notice that the measure of the subset $\mathcal{S} \subset \mathcal{W}$ of self avoiding walks is *exponentially small* compared to the measure of all maps (walks) •—•—•—... $\to \mathbb{R}^3$, that is

$$mes(\mathcal{S}_N) \sim (1-\delta)^N mes(\mathcal{W}_N) \text{ for } N \to \infty,$$

[7]This space is more complicated for real proteins, see lecture ???

but this is not properly represented by the above drawing.

Also notice that the *diameter* of $\mathcal{S}_{\mathcal{N}}$ is significantly *greater* than that of the ambient $\mathcal{W}_N \supset \mathcal{S}_{\mathcal{N}}$, where this diameter is measured for the distance in $\mathcal{S}$ defined by the lengths of shortest paths between points $S_1, S_2 \in \mathcal{S}$.

Probably, this diameter is infinite for large N, i.e. the space $\mathcal{S}_N$ of embeddings of strings of beads •—•—•... to the 3-space may be disconnected due to the possible presence of of so densely packed strings of or their fragments that they can't be unfolded.

If this happens, then $\mathcal{S}_{\mathcal{N}}$ is divided into exponentially many connected components, each carrying an exponentially small part of the full measure of $\mathcal{S}_N$, where the largest component is the one of the "fully unfolded" string $(001, 002, ...00N)$.[8]

Also there should be strings S, at least for special values of ε – the radius of the beads, certain parts of which being so densely packed/folded that they are unmovable. Then the connected components of such S would display all kind of dimensions in the range between d and D_N, where $d = d(\varepsilon)$ is a constant and $D_N = dim(\mathcal{S}) = 2N + 1$.

Exercise. Prove or disprove the above statements.[9]

Besides the geometry of $\mathcal{S}$ per se , we want to understand the structures of natural functions/observables on $\mathcal{S}$, such, for instance, as

$$W \mapsto diameter(S(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!- \ ... -\!\!\bullet)).$$

We shall return to all this in lecture??? but now let us take a glimpse on spaces of natural proteins sequences and "biological observables" on them such as the shape of a folded protein regarded as a function on the amino acid sequence space 20^N.

1.3 Suggested Reading

1. *Protein Physics*, A Course of Lectures by Alexei Finkelstein Oleg Ptitsyn (2nd Edition).

2. *Crystals, Proteins, Stability and Isoperimetry*, M. Gromov.
`https://cims.nyu.edu/~gromov/`

3. *The Science and Technology Behind the Human Genome Project* by Charles R. Cantor and Cassandra L. Smith.

4. *Unraveling DNA: The Most Important Molecule Of Life*, by Maxim Frank Kamenetskii (Revised And Updated Edition).

5. *The Logic of Chance: The Nature and Origin of Biological Evolution*, by Eugene V. Koonin

6. *The Plausibility of Life Resolving Darwin's Dilemma* by Marc W. Kirschner and John C. Gerhart.

7. Cell Biology by the Numbers, by Ron Milo, Rob Phillips.

8. *Lateral DNA Transfer: Mechanisms and Consequences*, by Frederic Bushman.

9. *Self-Avoiding Walk* by Gordon Slade,

[8] The distance between an S in this component and $(001, 002, ...00N)$ may serve as an upper bound on the folding time of S in some model.

[9] I didn't try to solve this exercise, but it seems not very difficult.

https://www.imsc.res.in/~sitabhra/teaching/asm17/Slade_MathIntelligencer_1994_Self_avoiding_walks.pdf

10. *The Self-Avoiding Walk* :A Brief Survey by Gordon Slade,
https://www.math.ubc.ca/~slade/spa_proceedings.pdf

11. *Lectures on Self-Avoiding Walks* by Roland Bauerschmidt, Hugo Duminil-Copin, Jesse Goodman, and Gordon Slade.
https://www.ihes.fr/~duminil/publi/saw_lecture_notes.pdf

12. *Self-avoiding walk, spin systems and renormalization* by Gordon Slade.
https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2018.0549