Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Biology, Geometry and Information

by

Jürgen Jost

Preprint no.: 39 2019

Biology, Geometry and Information

Jürgen Jost*

1 Introduction

This is an essay about the conceptual foundations of modern biology and the role that mathematics can play for biology.

Traditionally, two aspects have been considered as fundamental for or constitutive of life, *reproduction* and *metabolism*. The concept of *evolution* puts the emphasis on the first of them, reproduction. Some modern versions, like the notion of the selfish gene, go well with the general public, but fall short of capturing the complexity of life. An important property of biological reproduction is the transmission of *information*, rather than of material structures. Other approaches, like *autopoiesis* or *dynamics far from thermodynamical equilibrium*, put more emphasis on the second aspect, metabolism, that is, maintaining a biological organism and preventing it from disintegrating. Metabolism needs a constant inflow of matter and energy, not just of information.

In this article, I wish to develop a conceptualization that combines and intertwines the two aspects. I shall propose that the key feature of biological life is the *control and regulation of processes*. This can happen in a hierarchical or a reciprocal manner. The basic processes themselves are material and occur in time and space, *3-dimensional space* in fact. The latter will assign a more fundamental role to geometry than usually allowed for in theoretical biology. The control and regulation of processes, while possibly depending on material interventions, requires information, about which processes to select and how to control them, so as to satisfy the needs of the controller to build up and maintain its structure. Importantly, the controller thereby externalizes much of its requirements, and makes itself dependent on complex other processes in its environment. Complex life can only survive in a complex environment. In many regards, that environment has to be more complex than the controller itself. An extreme biological example are virusses that are entities consisting of a simple mechanism to control possibly very complex organisms for the purpose of their own proliferation. But control can also simply consist in the utilization of basic physical laws, like gravity, or properties of 3-dimensional space. The general principle is that what can be provided for by the physical, chemical, biological or perhaps social environment need not be manufactured by the system itself.

*Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, and Santa Fe Institute for the Sciences of Complexity, Santa Fe, New Mexico, USA

Since this principle can be iterated in the biological and social realm, ever more complex structures can build up in a hierarchical manner or may depend on each other in a reciprocal manner. Externalization by substituting the shaping and exploitation of external processes for internal ones, and internalization by the tighter control of originally independent processes then go hand in hand, as also emphasized in [16]. In fact, in [16] this is described as the interplay between niche construction and regulatory networks.

This paper is an extended version of my presentation at the Conference "Geometry and Phenomenology of the Living. Limits and possibilities of mathematization, complexity and individuation". I thank Luciano Boi, Carlos Lobo and Giuseppe Longo for organizing a very stimulating meeting. I also thank Klaus Scherrer for an inspiring collaboration on the conceptual foundations of molecular biology, and Manfred Laubichler and many others for stimulating discussions.

2 Biology and mathematics

Geometry, information, dynamics etc., that is, fundamental concepts that emerged in the introduction, are mathematical concepts. We shall consequently systematically draw upon mathematical thinking. To put this into perspective, let us first discuss the general question of what mathematics can contribute to biology. Some general possibilities are

- Methods for detecting structure in data
- Dynamical models of biophysical processes and their analysis
- Abstract conceptual analysis

Obviously, these approaches operate at very different levels. Therefore, let us consider some possible mathematical approaches in more detailed terms. We shall then see that in some sense, they cut across the different levels of the preceding list.

1. Information theory
2. (3-dimensional) geometry
3. Biophysical models and dynamical systems
4. Network analysis and generalizations, like simplicial complexes or hypergraphs

Let us discuss them in some more detail.

1. Information theory: Here the biologically fundamental question becomes what information is relevant. That is, information acquires its value only insofar it is related to the biological entity in question, by guiding its survival, maintenance or reproduction.

2. (3-dimensional) geometry: The fact that life occurs in 3-dimensional space is important, but often not addressed at all in theoretical biology. 3D enables, facilitates and/or constrains biological processes. Fewer dimensions would offer too few possibilities for spatial arrangement or interaction, while more dimensions might not constrain biological processes enough to prevent them from disintegrating.

3. Biophysical models and dynamical systems: Here, an important question is about the appropriate level of detail of the biological models. In fact, more detailed models sometimes yield less accurate or robust predictions than coarser ones. Some of this may simply be cases of overfitting, but the deeper reasons are not yet systematically understood from a biological perspective.

4. Network analysis and generalizations, like simplicial complexes or hypergraphs: Here, an important point that is quite generally ignored in the analysis of biological and other networks is that the edges express relations, and they, in place of vertices, should therefore be the basic objects of network analysis. In particular, the quantities utilized in network analysis should assign values to edges, rather than to vertices.

3 Three-dimensional geometry

Biological structure exist and interact in space. Space is 3D, although from the perspective of biological organisms not Euclidean, because gravity acts in one direction on the surface of the earth, and therefore, the degrees of mobility of and interactions between terrestrial organisms are often constrained to something more like 2D. On the scale of cells, this plays a more minor role. Nevertheless, one of the most important structures, the DNA is arranged in a one-dimensional manner. Why is this so? It is not 3D, because the interior of a 3D object is not accessible for readouts or copying. It is not 2D, because a linear structure is better adapted to sequential, temporal processing. That in turn is needed because there are bottlenecks like the ribosome where polypeptides are assembled. Likewise, replication seems to be less error prone and more energy efficient when arranged in a sequential, temporal manner instead of taking place simultaneously, like in a Xerox machine. When replication is carried out sequentially, the same copying molecules and structures can be used repeatedly. Of course, both the DNA itself and also its products, the polypeptides then acquire a 3D shape. For the polypeptides that constitute proteins, this is essential for their biological functions, because in that way particular motives can be exposed to interactions with other substances or shielded from such interactions. For the DNA, the spatial arrangement is important for the regulation of gene transcription, as emphasized by L.Boi [2]. For instance, via a suitable spatial organization, genomic regions that should be simultaneously transcribed can be brought into spatial proximity, even if their intrinsic linear distance on the DNA could be quite large, thereby facilitating coregulation, as originally proposed by

Képès and Vaillant [13], although their original solenoid model may have been too simple. The spatial organization of the DNA is certainly not as erratic as it originally looked, nor as regular as proposed in the first models, but very carefully orchestrated by specific proteins.

For the RNA, which is not only an intermediate between the DNA and proteins, but the crucial instance for regulation and processing (as well as for a host of catalytic roles), the secondary structure is important, achieved by pairwise bonds between complementary nucleotides in a linear sequence. Much of the processing is regulated by interactions with specific proteins, and in [11], a combinatorial code has been proposed. In turn, RNA molecules can also function as scaffolds for bringing specific groups of proteins together to induce their functional interaction. For that role, a 2D structure seems to be the most appropriate.

A caricature might then say that we proceed from the 1D DNA (information storage) via the 2D RNA (regulation and processing) to the 3D proteome (cellular function). Of course, as discussed, the DNA is organized in 3D, and RNAs and proteins not only have a 3D shape, but also interact in 3D.

The latter point indicates that 3D geometry is important not only for single structures, like proteins or the DNA, but also for the interaction between structures. It facilitates and constrains interactions at the same time. In 3D, substances can find each other more easily than in higher dimensions, but there are also constraints for simultaneous interactions. If some substance occupies a place in space, that place is no longer accessible to others. The effects may not always be easy to access. Let us consider the example of the toponome project of Walter Schubert, Andreas Dress and their collaborators [22]. By repeated staining and bleaching of a cell slice, they can record the positions of about 100 proteins in that slice. In particular, one then has data about the colocalization of proteins. These data can be arranged in a simplicial complex. The vertices of that simplicial complex stand for the various proteins. Two vertices are connected by an edge if the two corresponding proteins frequently occur in neighboring positions. Here, one can set some threshold, how often those proteins should occur together in order to speak of cooccurence and introduce the corresponding edge. Similarly, one inserts a two-dimensional simplex, that is, a triangle with three vertices, when all three corresponding proteins frequently occur together, and not only each pair among them. And similarly for higher dimensional simplices. Proteins can interact only when they are in spatial proximity, that is, when they cooccur, and so, this simplicial complex represents some kind of geometric backbone for the interaction patterns. Of course, interactions are realized by chemical affinities. This in turn leads to the question which of those potential chemical reactions are actually realized in the cell. 3D geometry may prevent certain chemically possible interactions from happening, because not all them can happen simultaneously in space. The mathematical question then is what constraints this creates for the topology of the simplicial

complex whose construction we have just described. To study such a simplicial comples, Betti numbers (dimensions of homology groups in algebraic topology, see for instance [7, 10]) and geometric invariants like Laplacian spectra [5] can be used for qualitative comparison of colocalization patterns in different cells (e.g. healthy vs. diseased).

At another scale, the organization of the brain is also three-dimensional. Since not every structure can be in spatial proximity with every other structures, more distant structures need to be connected by biological wires or cables. Sending information through such cables takes time, and this then slows down the processing speed for signals entering the brain. Making the cables thicker increases the speed, but then fewer such cables can fit into some given region. Therefore, there is an optimization problem for the arrangement of the various cortical and subcortical structures and the wiring between them, so that the most important signals can be processed as fast as possible. But since those important signals and the adequate responses to them may be quite heterogeneous, compromises between the processing efficiencies of various data are necessary. We may then ask how good the solution is that biological evolution has found for the spatial organization of the brain, or whether another, perhaps radically different or more systematic design might be superior for the problems that the brains of current organisms have to handle.

A closed surface can shield its interior from external perturbations or influences. This inaccessibility has positive and negative aspects. An obvious positive aspect, emphasized for instance in Maturana's and Varela's theory of autopoiesis, is that a cell wall prevents the cell from disintegrating and at the same time, being selectively permeable, enables the inflow of needed material. But then also interactions with external substances that should not or cannot be admitted into the cell need to be mediated by receptors on the cell wall and internal signalling cascades.

And we also recall that inside the cell, the DNA could not be intrinsically three-dimensional, as otherwise it would not be accessible for transcription and replication.

We conclude that information, regulation and geometric structure are interwoven, and each theoretical treatment should keep that in mind.

4 Interactions, networks and hypergraphs

Colocalization patterns of proteins constitute temporal snapshots. They constitute preconditions or show results of metabolic or other biochemical reactions. In those reactions, also other substances are involved, and the proteins may catalyze those reactions. These reaction sets are properly modelled not as simplicial

complexes, but as *chemical hypergraphs*. These are structures where two sets of vertices, standing for educts (ingredients) and products of chemical reactions, are connected by hyperedges, standing for the chemical reactions. These sets of vertices need not be disjoint, as catalysts should be counted as both educts and products of reactions. The formal analyis of such chemical hypergraphs has been started in [12]. Such hypergraphs can be analyzed via Laplacian spectra or by the distribution of metric curvatures. Chemical reaction networks are constrained by stoichiometry. In this regard, a theory has been developed by St.Schuster and others for decomposing metabolic pathways into elementary modes, see for instance [24, 23, 21, 14]. The availability of external ingredients and energy (provided by ATP) and reaction rates, but also spatial organization, constrain how much can be produced in parallel or sequentially. Again, coordination, regulation, and control are necessary.

5 Regulation

As we have already seen, the coordination of processes can be achieved in principle by spatial proximity (geometry), or by joint signals (information) These are not alternatives, but can be flexibly combined.

And spatial interactions may have a dual role. We again recall the example of the interaction of RNAs and proteins (RNPs). There are two possible functional roles:

1. The RNA serves as a scaffold for protein interactions. Thus, a specific spatial organization of the cell can guide the specificity of interactions, or

2. the regulation of gene expressions via a combinatorial code, for the coordination of expressions of specific collections of genes.

The first item emphasizes again the role of topology. Klaus Scherrer and myself have there proposed the term *topon* for a geometric configuration of regulatory significance. The second item is systematically developed in [18, 19, 20]. An important biological principle is (pre-)mRNA only further processed when some binding proteins are removed. That is, the removal of individual proteins shared by a specific collection of mRNAs enables the coordinated activation of specific sets of genes. . Here, we see the power of combinatorics [11]. The binding motifs for those proteins are contained in the RNA sequence, and so, one and the same stretch of RNA may have both a coding and a regulatory role, and we have proposed the term *genon* for such a regulatory motif superimposed on a coding sequence. An mRNA has about 20 such binding sites for proteins, each of them shared with some other RNAs. Thus, taking for instance 5 binding sites, there is a specific group of mRNAs that have all of them in common. When all binding sites are occupied by their corresponding proteins, the mRNA is not further processed, but sits there in some kind of dormant state. When, however, a certain number, say 5, of those proteins is removed, the mRNA is

further processed and translated into a polypeptide. Thus, when some signal removes those 5 proteins from all their binding sites, a specific group of mRNAs is translated. In other words, we have some combinatorial scheme that enables the cell to translate a specific set of genes, according to specific requirements. The numbers involved, of different binding sites, of binding sites per mRNA and of proteins to be removed for processing, are such that there is a huge number of combinatorial possibilities. See [11] or [6] for details.

6 A fundamental thesis

We shall now formulate our fundamental thesis (see also [8, 9] for different contexts) and explore its consequences.

Thesis 1. *The key principle of biology is that a process can control and regulate other processes.*

Examples:

- Promoter, repressor etc. sites at the DNA are unspecific for the coding regions, but reflect the regulation schemes
- There are many general combinatorial regulatory mechanisms at RNA level (interactions between different RNAs or RNAs and proteins)
- Hoxgenes as general regulatory mechanisms across species [4]
- Principle of allosteric inhibition (Monod, Changeux, Jacob [17])
- Insects have a general, unspecific control mechanism for transforming sensory input in motor activity. They can therefore flexibly couple sensors to actuators.
- In the research direction of Evo-Devo (which can be seen as a challenge to the Neodarwinian paradigm), the key is the reorganization of control mechanisms (see for instance [3, 15])

These examples suggest our next thesis.

Thesis 2. *The content of these controlled processes matters only insofar as it serves the controlling process.*

This then has implication for the question "*What is relevant information?*". It leads to a new concept of biological information.

Thesis 3. *Relevant information is only what is needed for regulation and control. This may be very little.*

Let us discuss some examples and applications.

1. A virus, in order to start with perhaps the most extreme example, only needs to "know" how to find a host and inject its DNA or RNA into that hosts cells. Therefore, the genetic information of the virus can be very short. The virus controls the host's metabolic processes to ensure its own replication. How those processes operate is irrelevant.

2. A higher animal, a mammal for instance, has the evolutionary choice about which metabolic products to manufacture itself and which to simply take in as food. Vitamins are a good example. They are essential for the metabolism, but their production is externalized. Thus, the animal no longer needs to store the information about the necessary metabolic processes in its genome, but rather the information how to acquire food containing the necessary vitamins. Thus, the production is externalized.

3. A common aspect of the two preceding examples is that a biological organism or process (if we may consider the replication of a virus as a biological process) depends on an environment that may be vastly more complex than itself. The metabolic information about how to produce vitamins may be much higher than the information about finding the appropriate food source, but only the latter is needed for the organism or process.

4. Biological organisms not only exploit other organisms or processes in their environment, but also, and perhaps even more basically, physical laws and regularities. For instance, gravity is actively exploited in much of animal locomotion. Our bodies are adapted for walking in the presence of gravity of a very particular strength. Robotics has recently learned to also utilize the forces of gravity, instead of carefully programming the positions of all the joints of a walking robot. That is called *embodiment*.

5. As a consequence of the principle that a biological organism depends on both a complex environment and on the operation of physical laws, it is doubtful whether we can ever establish human life on other planets. While we may be able to control other physical parameters like the temperature or the oxygen supply, our bodies are not adapted to operate under a different gravity strength. And whatever artificial biological environment we may be able to create, it may not be complex enough to sustain human life in the long term.

6. Ashby's law of requisite variety [1] is incorrect. That law says that a system needs to maintain enough variety to match all external perturbations if it is to persist in the presence of such perturbations. In fact, as a consequence of our theses, the system needs much less. It simply needs to control processes, either directly those that generate the perturbations, or others that handle those perturbations.

7. Most of the preceding examples and arguments present instances of externalization, that is, when external processes are created or utilized to perform some function for the organism in question. Many such examples

are instances of niche construction. As Laubichler and Renn [16] point out, the reverse is equally imported, where external processes are internalized. For instance, the mitochondria in eukaryotic cells derive from biological entities, essentially bacteria, that were originally independent, but then incorporated into those eukaryontic cells for metabolic processes. More generally, regulatory networks become ever more sophisticated to control ever more complex internal processes.

8. In the same direction, the answer to the question why the simulation of protein folding is so difficult is probably not physical (an energy landscape with many metastable states), but genuinely biological: The energy landscape evolved to provide flexibility to switch between different conformations.

Thus, we see that when viewed from the perspective of the above theses, many very diverse biological phenomena fall in place conceptually and acquire evolutionary significance.

References

[1] W. R. Ashby, An Introduction to Cybernetics, Chapmann & Hall, London, 1956

[2] L.Boi, Plasticity and complexity in biology: Topological organization, regulatory protein networks, and mechanisms of genetic expression. In: G.Terzis u. R.Arp (eds.), Information and living systems, MIT Press, 2011, pp.205–250

[3] Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to Diversity (2nd ed). Malden, Mass.: Blackwell Science.

[4] W.Gehring, Master control genes in development and evolution, Yale Univ.Press, 1998

[5] Horak, D., und Jost, J. (2013). Spectra of combinatorial Laplace operators on simplicial complexes. Advances in Mathematics, 244, 303-336.

[6] J.Jost, Mathematical methods in biology and neurobiology, Springer, 2014

[7] J.Jost, Mathematical concepts, Springer, 2015

[8] J.Jost, Leibniz und die moderne Naturwissenschaft, Monograph, Series *Wissenschaft und Philosophie, Science and Philosophy, Sciences et Philosophie*, Springer, in press

[9] J.Jost, Biologie und Mathematik, Monograph, Springer, in press

[10] J.Jost, Mathematical Principles of Topological and Geometric Data Analysis, Monograph, in preparation

[11] J.Jost, K.Scherrer, Information theory, gene expression, and combinatorial regulation - A quantitative analysis, Theory Biosci. 133, 1–21, 2014

[12] J.Jost und R.Mulas, Hypergraph Laplace Operators for Chemical Reaction Networks, arXiv:1804.01474.

[13] F. Képès, C. Vaillant. Transcription-Based Solenoidal Model of Chromosomes. Complexus 1, 171–180, 2003

[14] S.Klamt, J.Stelling, Two approaches for metabolic pathway analysis?, TRENDS in Biotechnology 21, 64–69, 2003

[15] M.D. Laubichler (2007): Evolutionary Developmental Biology, in David Hull and Michael Ruse (eds.) Cambridge Companion to the Philosophy of Biology, Cambridge University Press, pp: 342–360

[16] M.D. Laubichler and J. Renn, Extended evolution: A conceptual framework for integrating regulatory networks and niche construction. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 324(7):565–577, 2015

[17] Monod, J., Changeux, J. P., und Jacob, F. (1963). Allosteric proteins and cellular control systems. Journal of Molecular Biology, 6(4), 306–329.

[18] K. Scherrer, J. Jost, The Gene and the Genon Concept: A Functional and Information-theoretic Analysis, Molecular Systems Biology 3 (87), 2007

[19] K. Scherrer, J. Jost, Gene and genon concept: Coding versus regulation, Theory Biosci.126, 65–113, 2007

[20] K. Scherrer, J. Jost, Response to commentaries on our paper *Gene and genon concept: coding versus regulation*, Theory Biosc.128, 171–177, 2009

[21] C. Schilling, S. Schuster, B. Palsson, R. Heinrich, Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol. Prog. 15, 296–303, 1999

[22] W. Schubert, B. Bonnekoh, A. Pommer, L. Philipsen, R. Bockelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A. Dress. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nature Biotech. 24, 1270-1278, 2006

[23] S.Schuster, D.Fell, T.Dandekar, A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks, Nature Biotech. 18, 326–332, 2000

[24] S. Schuster, C.Hilgetag, On elementary flux modes in biochemical reaction systems at steady state. J. Biol. Syst. 2, 165 182, 1994