Rend. Mat. Appl. (7).
Volume 40 (2019), 165 – 176

RENDICONTI DI MATEMATICA
E DELLE SUE APPLICAZIONI

The actual status of quantitative approaches in Biology: problems and perspectives

Alessandro Giuliani

Abstract. *Biomedical sciences are traversing a "reproducibility crisis": it is estimated that more than 85% of basic research results in cancer research [25] are not replicable, and similar figures hold for other investigation fields. The urgent need of re-thinking the quantitative approach to biology is evident, here I try to enucleate the roots of the problem and sketch some possible solutions with a particular emphasis on complex network approaches.*

The 2005 John Ioannidis paper "Why most published research findings are false" [14] was a real shock for the biomedical science community. After more than ten years, it is now evident Ioannidis unveiled a real information crisis plaguing biomedical sciences due to both the inadequacy of the great majority of scientists to grasp the real meaning of statistical approach [18] and to the positioning of biomedical research at an inadequate (too detailed) noise-dominated level of investigation [23].

The increasing importance that molecular biology gained in the last thirty years, made the majority of biologists to think the only 'relevant explanations' must be looked for at the molecular level, i.e. the paradigm of a biological explanation should be in the form 'gene A provokes phenomenon (disease, phenotypic trait ...) B'. The existence of a single 'explanatory layer' is in sharp contrast with what we know about complex structured systems, where a multi-layer causality is at work. Ecology (with no doubt the biology field with a most sensible use of quantitative tools) recognized since many years that the 'most microscopic' level of organization is not necessarily the place where 'the most relevant facts do happen'. On the contrary, the most fruitful scale of investigation (in the major part of the cases) is where 'non-trivial determinism is maximal' [19]. That is to say, the scale more 'rich' in meaningful correlations between features pertinent to micro- and macro- scale or, to use an ecological term: the mesoscopic realm.

2010 Mathematics Subject Classification: 92C42, 62P10, 92B15.
Keywords: complex networks, organized complexity, information crisis, big data.
© The Author(s) 2019. This article is an open access publication.

Non-trivial determinism can be defined in terms of prediction error as:

$$\text{Prediction} \quad r^2 = 1 - \frac{E^2}{S^2}$$

In the above formula, E is the mean prediction error and S the standard deviation. In the case of a simple linear regression in which a dependent variable Y must be predicted by an independent variable X, the non-trivial determinism is nothing else than the usual squared Pearson correlation between the two X and Y variables. The formula can be extended to any other situation in which we wish to predict a system feature Y located at a hierarchical higher layer with respect to X, moreover both X and Y do not need to represent single variables but any suitable set of information at any definition scale.

Consequently, in the 'many Y'/'many X' case, the non-trivial determinism corresponds to the first canonical coefficient [11] while in the case of a binary diagnosis to the area below the ROC curve [13].

It is worth focusing on the specification 'non-trivial' attached to the word determinism. The statement *'Any protein is made up of 20 different amino-acid residues'* exactly determines a shared feature of the chemical composition of the protein molecules but, for the same fact it holds for all the proteins in the same way, it gives no information on the differences among protein molecules. This is 'trivial' in the case of biology, where the relevance of a scientific statement stems from the ability of getting rid (e.g. by means of establishing a meaningful correlation) of the variance of the system at hand.

A famous joke, reported in [8], clarifies this point: a very rich man, very fond of horse races, hired a top-class mathematician (e.g. Kurt Gödel) and a physicist (e.g. Albert Einstein) to build a model enabling him to exactly predict the winner of any horse race. After one year, both scientists returned to the rich man with their results. Gödel said *'Sir, I cannot say which is the specific horse who will win the race, but I discovered the solution to the problem exists and it is unique'.* The sponsor of the research is not satisfied at all and asks Einstein if he can say something more useful, Albert says *'Why did you ask Kurt? You must know mathematicians have no sense of reality; on the contrary, I have the exact solution indicating the specific winner of the race. It applies only in the case of spherical horses but I am convinced this is not a problem.*

Beside the delusion of the rich man, the joke reports the two basic inconsistencies of the mathematical and physical way of reasoning when dealing with biology: the lack of interest of both too abstract solutions and of sketching ideal cases to approximate real world.

One the fathers of information science, Warren Weaver, in his fundamental "Science and Complexity" 1948 paper [26], proposed a three-class partition of science into: 1. Simplicity, 2. Disorganized complexity, and 3. Organized complexity.

The first class (Simplicity) refers to the classical use of quantitative methods in science. Class 1 problems allow for an extreme abstraction (e.g. a planet can

be thought as a dimensionless 'material point'). This allows generating differential equations predicting the behavior of the studied system while relying on the stability in both space and time of the experimental (observational in the case of astronomy) results. The drastic reduction of the relevant properties to take into consideration down to very few basic features like mass and distance, allows for a straightforward appreciation of classical physics. This incredibly successful strategy is only very rarely possible in biology (e.g. this is why a great part of biomathematics redounds around Volterra-Lotka prey-predator models in which the 'Gödel-answer' to the joke has important 'real world' consequences).

Problems of Disorganized complexity (class 2) allow for a greater generalization power than class 1 by means of a very different style of reasoning. Here, the predictive power stems from the generation of very coarse grain macroscopic descriptors corresponding to gross averages on a transfinite number of atomic elements. Thermodynamics is the brightest example of this style of reasoning: emergent collective parameters like temperature or pressure allow for an accurate control of system behavior without the need to go into microscopic (noise-dominated) details.

Both the approaches must fulfill very stringent constraints. Class 1 approach asks for few involved elements interacting in a stable way, class 2 style needs a very large number of identical particles with only negligible (or very stable and invariant) interactions among them. Biological systems, only in very few cases do fulfill these constraints, so we step into Weaver's third class (Organized complexity).

Organized complexity arises when many (even if not so many as in class 2) non-identical elements each other interact with time-varying correlation strength. This provokes an extreme context dependence of the results so giving rise to the 'information crisis' biology is now experiencing. This is the 'middle kingdom' where life sciences live that was recognized as the XXI century frontier of basic science [15].

Before going ahead, it is worth reporting the original figure of the Weaver paper sketching the three realms of science: Weaver claims that when dealing with complex organized systems, the focus of the investigation must shift from the detailed analysis of single elements to their wiring pattern.

The clarity of the Weaver's message faded away by the action of a drastic terminological (and philosophical) revolution: it is not without consequences referring to quantitative approaches by the term 'Informatics' instead of 'Mathematics' (as actually was the case in the great majority of biological applications).

Bioinformatics revolution started with the need of generating (and storing) very long symbolic strings correspondent to the sequences of biopolymers (DNA, RNA and protein molecules). The analysis of symbolic strings is probably the 'most classical' problem of informatics dating back to the very birth of the discipline since Alan Turing seminal studies [24]. This act-of-birth influenced the development of the relation between bioinformatics and biological sciences. Bioinformatics tools are considered as purely technical devices (like a fridge or a spectrophotometer) helping the biologist to answer questions that arise from (largely

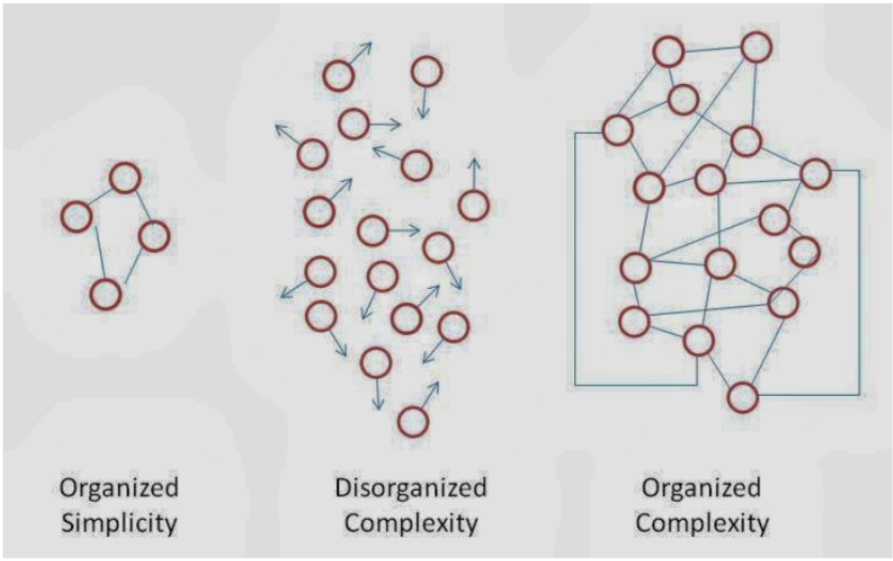

Figure 1: Circles represent the elementary players, the lines their mutual relations. The lines of the graph in the middle (disorganized complexity) correspond to the trajectories of the particles whose interactions are both random and contingent. The third panel depicts organized complexity: the interactions are both non-negligible and time varying.

qualitative) speculation. The 'quantitative step' is very stereotyped and reduces to a 'pattern matching' in which the problem is to find the 'maximal superposition' between a 'query' (e.g. an unknown biopolymer sequence found in a sample) and a 'target' (sequences of proteins whose physiological role is known).

This state of affairs transforms the informatician into a 'servant' solving a practical problem with no relevant role in the emergence of new insights. The generation and interrogation of static (only growing for brutal addition of new data) repositories where to look for potentially useful hints for the problem at hand, constitutes the almost totality of the work. This approach, in the last two decades, enlarged its range from biopolymers to gene expression (transcriptomics), metabolic pathways (metabolomics), medical diagnoses but substantially stays well inside the 'pattern matching' class of problems. Going back to the horse-race joke, this ends up into another inconsistency trap: producing an a posteriori solution tailored on each single race with no possibility to be generalized to any future competition.

The maximization of 'non-trivial' determinism instead, asks for both the consideration of different races peculiarities (variance) and their 'explanation' (in terms of minimal prediction error) by means of a generalizable rule based upon other views of the system. These 'views' should be sufficiently meaningful to allow

for a common explanation of an entire 'class' of problems and not only limited to a specific instance.

The information crisis only exacerbated the 'Bioinformatics' stereotypy, as evident in nowadays 'Big Data' enchantment. The 'Big Data' proposal starts from a correct assumption: the nowadays information crisis is an 'overfitting' crisis: in presence of too many degrees of freedom (being they genes, proteins, metabolic reactions..) as consequence of the development of 'high-throughput' techniques allowing to measure thousands of different descriptors of (relatively few) independent observations, the risk of chance correlations became unbearably high [8, 12].

The 'Big Data' proposal to overcome this problem is (roughly): *"Let's give up with theory-driven experimentations and let's look, without preconceived ideas, to the 'whole-thing' (the development of various –omics makes this possible): the emerging correlations will allow for new ideas and findings spontaneously appear in a data-driven way"* [13].

Notwithstanding the increasing funding of 'Big Data' initiatives, it is sufficiently clear that the pure enumeration of single relevant correlations across a huge number of variables only exacerbates the reproducibility crisis [8]. Pure data-driven approaches set forth by the 'Big Data' extremists claiming for the 'end of scientific method' (see for example [1]), promise to become the Heaven of chance correlations (see [2] for a very interesting critic to the pure informatics approach to science).

What we really need is to look for 'Universal Organization Principles' of complex systems (Weaver Class 3) moving from the 'microscopic' (single nodes) to the 'mesoscopic' (wiring pattern) level [15]. In order to perform such a 'quantum leap' we could profitably make use of an information science concept different from sequence capturing the essence of organized complexity.

As pointed out by Nicosia et al. [17]: *"Networks are the fabric of complex systems"*. This is why different investigation fields –from protein science [6] to neuroscience [12]– build upon the consideration that shared organization rules should give rise to similar phenomenology, independently of the nature of the constituting elements. The quest for 'network laws' largely independent of the nature of the constituting nodes of the network, stems from the work of the Dutch electrical engineer Bernard Tellegen [21] that developed a sort of conservation principle of both potential and flux across a network analogous to Kirchoff's laws. The flux does not need to be an electrical current and the same holds for the potential. Any system modeled by a set of nodes linked by edges (being them metabolites linked by chemical reactions transforming one into the other or mutually interacting persons in an office...) has similar emerging properties independently of the physical nature of nodes and edges. As aptly stressed in [16], the theorem opens the way to a sort of 'network thermodynamics', whose principles are strictly dependent from wiring architecture while largely independent of the constitutive laws governing the single elements.

Formalizing a given problem in terms of a graph (a mathematical graph is

equivalent to a network expressed in terms of its adjacency matrix) allows for a thermodynamic-like approach (here focusing on relations and no more on means like in Weaver class 2) to be applied to complex systems.

We can roughly describe the network approach as the answer to the question *"What can we derive from the sole knowledge of the wiring diagram of a system?"*

An adjacency matrix (and consequently a complex network) can generate from any sensible correlation metrics applied to the elements of a system. A correlation matrix reporting the pairwise Pearson coefficients between continuous variables, Euclidean distances computed at any dimensionality between discrete landmarks (e.g. amino-acid residues location in 3D protein structure, species abundance profiles) or the phase coherence of electrophysiological signals are only some examples of the situations that can profitably expressed as graphs.

The most basic level of quantitative description of graphs (correspondent to basic descriptive statistics of random variables) is the computation of so called 'graph invariants' [3]. These invariants are relative to local (single nodes), global (entire network), and mesoscopic (clusters of nodes, optimal paths) levels.

Thus, the "degree" (how many links are attached to a given node) is a local descriptor, the "average shortest path", corresponding to the average length of minimal paths connecting all the node pairs, is a mesoscopic feature, while the general connectivity of the network (density of links) is a global property [3].

Figure 2 reports an exemplar network structure with the indication of some relevant descriptors of the wiring architecture: the values of local, mesoscopic and global descriptors are naturally linked to each other by the peculiar network architecture. This fact creates a 'natural' microscopic-macroscopic link devoid of any strong theoretical assumption.

This very basic level of description allows for deriving useful biological information: protein structures can be formalized as graphs (protein contact networks) having amino-acid residues as nodes. A link is established between two i and j residues if $d(i,j) < R$, where $d(i,j)$ is the Euclidean distance between i and j and R corresponds to Van der Walls radius, the maximal distance the two residues can engage an effective relation (i.e. the maximal distance they can be considered in contact) [6]. Fig.3 reports a protein contact network analysis of hemoglobin highlighting different graph-based representation of the same molecule [4].

Hemoglobin is made of four sub-units (two alpha and two beta subunits pairwise identical) coded with different colors in panel A. Panel A representation builds upon the so-called 'secondary structure', i.e. the relative arrangement of neighboring residues along the chain. Chemically speaking, a protein is a polymer where the monomers (residues) are covalently linked one after the other to form a continuous chain or sequence, the chain, when folds in three dimension space, presents three main patterns (secondary structure) accounting for the local arrangement of adjacent residues along the sequence. These patterns are named: 'helix', 'beta-sheet' and 'random coil', here the great part of the chain is arranged in terms of 'helix' pattern. Panel C reports the corresponding protein contact network: the dots are the single residues while the contacts between them are the blue segments.

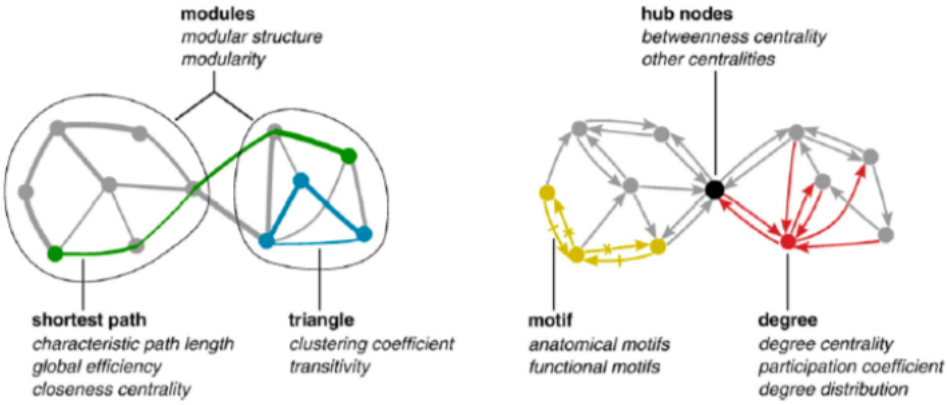

Figure 2: Modules correspond to subset of nodes having much more links among them than with other nodes of the network. Measures of centrality (closeness, betweeness...) describe nodes in terms of the number of shortest paths traversing them. Shortest path is the characteristic metrics for networks: they correspond to the shortest distances (in terms of number of nodes/links to be traversed) for linking pairs of nodes.

The color of the dots correspond to the role they have in the network: white dots (R1) engage links almost exclusively with residues within their 'module'.

This situation is clarified in Panel B, where the points correspond to the amino-acid residues that are projected into a bi-dimensional space having as ordinate the within-module degree (a normalized score proportional to the number of within-module contacts) and as abscissa the Participation coefficient, an index proportional to the number of extra-module contacts. The bi-dimensional plane is subdivided into four regions (R1-R4) that correspond to different topological roles exerted by the nodes in the network. These roles, defined by Roger Guimerà and Luis Amaral in 2005 [9], define a cartography that can be applied to any network, despite the nature of their nodes and links according to network thermodynamics principles [16]. The cartography classes are as follows (Table 1):

Hub character (Module)	Regions	Within-Module z score	Participation coefficient
Non-hub	R1: ultra-peripheral	$z < 2.5$	$P < 0.05$
Non-hub	R2: peripheral node	$z < 2.5$	$0.05 < P < 0.625$
Non-hub	R3: Non-hub connector	$z < 2.5$	$0.0625 < P < 0.8$
Non-hub	R4: Non-hub kinless node	$z < 2.5$	$P > 0.8$
Hub	R5: Provincial hub	$z > 2.5$	$P < 0.3$
Hub	R6: Connector hub	$z > 2.5$	$0.3 < P < 0.75$
Hub	R7: Kinless hub	$z > 2.5$	$P > 0.75$

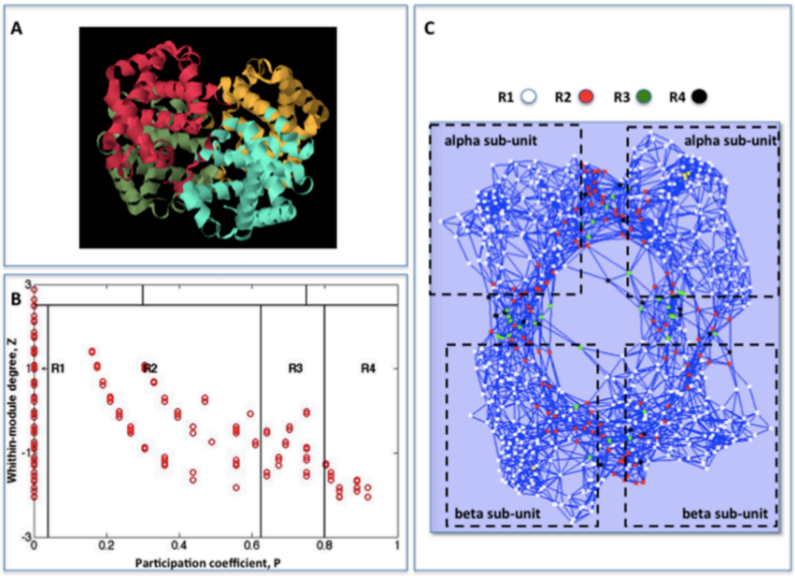

Figure 3: Protein network exploitation of Hemoglobin protein molecule. The alpha and beta subunits are super-imposed to protein contact network (panel B): their perfect correspondence with modules emerging by the network wiring is a proof-of-concept of the fact network formalism caught the essential of hemoglobin structure. Red, Green and Purple dots correspond to nodes with an increasing 'Participation coefficient', i.e. to nodes engaged in links with an increasing 'extra module' contacts (R2 to R4).

The term 'hub' comes from air traffic and originally denoted an airport with an exceedingly high number of flights departing from (or arriving to) it, in complex network jargon, a 'hub' is a node with an extremely high degree with respect to others. In the above cartography, the 'hub' character of a node is intended only in relation to its module (this is why it depends only by its within-module z-score). The classification of nodes into the seven cartography classes (R1-R7, see Table 1) directly stems from the relative proportion of within-module and extra-module (participation coefficient) contacts. The ranges defining the seven classes were assessed by a statistical physics approach based on a huge number of simulations in which thermodynamic properties of the network were computed (melting point, percolation threshold, stability, see [10]) after perturbing the network in different regions.

The first step to apply the above cartography is the separation of the network into modules, this partition is usually accomplished by means of spectral methods [4] directly applied on the adjacency matrix of the graph. If we consider the above sketched approach as applied to a protein molecule (in this case hemoglobin), it is immediate to note how we rely upon a drastically reduced set of information

with respect to the full rank protein structure information present in the three-dimensional coordinates of the amino-acid residues. First of all the amino-acid residues are considered as identical while in fact they are very different as for their physico-chemical properties. This is totally in line with Weaver class 3 problems: the focus is only on the wiring structure of the system, the elements differ among them only as for their role in the wiring diagram. The actual distances between the residues are substituted with a binary classification: 'contact' vs. 'non-contact' and the sequence information (order along the chain) as well as the secondary structure patterns are not taken into consideration.

Notwithstanding that, we are able to exactly recover the actual sub-units of the molecules that perfectly match the modules computed by the adjacency matrix (Fig.3, panel C), while the residues in charge of between sub-units contacts are recognized by their position in the cartography plane (nodes in R3 and R4 regions). It is worth noting that all the protein molecules, notwithstanding their huge differences in both form and physiological function, share some strong similarities in their nodes distribution in terms of cartography (e.g absence of hubs, very similar relative frequencies of R1-R4 classes). Thanks to the possibility to link the regional classification to network thermodynamics [9], this allows to derive general properties of the proteins as their allosteric character (ability to transfer a signal across the molecule without any loss of information [7]) while in the same way to locate the most functionally relevant residues [20].

It is virtually impossible to derive the above results keeping into consideration the full-rank information of the protein structure; this happens because the full-rank information is 'flat': there is no possibility to a priori discriminate between relevant from non-relevant between residues distances, there is no possibility to attach a 'protein-level meaning' to the chemical diversity of the amino-acid residues. In the same way an all-encompassing 'super-model' taking into consideration the whole information is bound to failure for both overfitting and lack of generalization.

This is immediately clear by comparing the impossibility to discriminate between 'allosteric' and 'non-allosteric' proteins by the consideration of the full rank information (relative distances between configurations on the basis of 3D coordinates of residues), while this difference can be grasped in terms of mesoscopic network descriptors (shift or amino-acid residues across cartography regions) [6]. This feature has to do with the recognized superposition of 'sloppy' and 'stiff' parameters of a given model [23]: focusing on the 'stiff' part of information is productive, considering on the same level 'sloppy' and 'stiff' features is not productive.

This is exactly the point from where we started: the information crisis affecting the biomedical sciences stems from the accumulation of a plethora of tiny details without any possibility to discriminate between the relevant and irrelevant ones while the 'Big Data' brutal proposal 'Keep them all' does not work because of the consequent overfitting and chance correlation deluge. The crucial role of theory in physical sciences was exactly to suggest to experimentalists where to look for

getting rid of a given problem without being lost in a plethora of irrelevant features. This is why is sufficient to concentrate on viscosity and density to study different liquids hydrodynamics, despite their microscopic diversity [19], on the other hand, the failure of a pre-established theoretical suggestion to cope with a given phenomenon is the starting point for sketching a new theoretical frame. Beside the 'organized complexity' peculiarities described by Weaver [26], there are many other reasons to take into consideration for explaining the virtual lack of operationally effective theories in biology [5], what is for sure is that, until now, the classical path to scientific progress typical of 'hard' sciences largely failed in biology. Biomedical sciences progress resided on the force of 'intuition' that allowed particularly gifted scientists to identify crucial phenomena endowed with a generalization power (e.g. ecological niche, immune response, energy metabolism, polymers carrying biological relevant information in their sequence, differentiation paths, infective agents ...).

Now we need something different, and network-thermodynamics [16] allowing for the generation of a statistical mechanics approach only based on the correlation structures of the studied systems is a very promising avenue for a fresh new start. The success of this style of reasoning strictly depends upon a sensible segmentation of the systems into parts. While is almost immediate to parse an artificial system into their constituting elements (think of an electrical circuit or an engine) this is not the case when dealing with natural systems where scale effects exert a crucial role and make the difference between a good and a bad partition [19]. This implies that a mathematician involved in a biomedical research enterprise can by no means be considered as an 'expert-in-numbers' to be called only at the very beginning (sketch of the experimental plan) and at the very end (data analysis) of the process. On the contrary, the mathematician must interact with biologists all along the entire research in the development of the most apt formalization of the system at hand. This makes necessary a deep cultural change involving both biological and mathematical sides of the coin. The mathematicians must make an effort to go in depth into the nature of the problem without being only concerned with 'rigor' and 'abstraction' on the other hand the biologists must push themselves toward a the need to translate their ideas into a formal and rigorous alphabet learning acquiring the some abstraction ability. Both the academic world and funding policies do not foster such a convergence when the overcoming of hyper-specialization is essential in order to solve the actual crisis [25].

There is an urgent need that statements like *'Drug A provokes a drastic decease of average shortest path of protein contact network'* can be accepted as a meaningful explanation [3] without the need to go in depth into a specific amino-acid residue or binding site. In any case, times are rapidly changing, it is sufficient to interrogate a scientific literature repository with the statement "complex networks AND gene expression" (3.020.000 results in Google Scholar the 21st January 2019) to get the sense that biomedical sciences are actively re-shaping, this process will have deep cultural and practical consequences. There is plenty of work for mathematicians, given they dare to (partially) abandon a too extreme rigor and start

to take interest in the 'real biological content' of the studied models.

References

[1] Anderson, C.: The end of theory: The data deluge makes the scientific method obsolete. Wired magazine **16**:7 (2008)
[2] Calude, C.S., Longo, G.: The deluge of spurious correlations in big data. Found. Sci. **22**:3, 595–612 (2017)
[3] Csermely, P., Korcsmáros, T., Kiss, H.J. , London, G., Nussinov, G.: Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol. Ther. **138**, 333–408 (2013)
[4] Cumbo, F. , Paci, P., Santoni, D., Di Paola, L., Giuliani, A.: GIANT: a cytoscape plugin for modular networks. PLOS ONE **9**:10, e105001 (2014)
[5] Dhar, P.K., Giuliani, A.: Laws of biology: why so few? Syst. Synth. Biol. **4**:1, 7–13 (2010)
[6] Di Paola, L., De Ruvo, M., P. Paci, P., Santoni, D., Giuliani, A.: Protein contact networks: an emerging paradigm in chemistry. Chem. Rev. **113**:3, 1598–1613 (2012)
[7] Di Paola, L., Giuliani, A.: Protein contact network topology: a natural language for allostery. Curr. Opin. Struct. Biol. **31**, 43–48 (2015)
[8] Giuliani, A., Zbilut, J.P.: The relevance of physical and mathematical modes of thought on complex systems behavior in biological systems. Complexity **3**:5, 23–24 (1998)
[9] Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**:7028, 895–900 (2005)
[10] Guimerà, R., Amaral, L.A.N.: Cartography of complex networks: modules and universal roles. J. Stat. Mech. **P02001**, 1–13 (2005)
[11] Härdle, W., Simar, L.: Canonical Correlation Analysis. In: Applied Multivariate Statistical Analysis, pp. 321–330, (2007)
[12] Hauser, T.U., Fiore, V.G., Moutoussis, M., Dolan, R.J.: Computational psychiatry of ADHD: neural gain impairments across Marrian levels of analysis. Trends Neurosci. **39**:2, 63–73 (2016)
[13] Heagerty, P., Zheng, Y.: Survival model predictive accuracy and ROC curves. Biometrics **61**:1, 92–105 (2005)
[14] Ioannidis, J.P.: Why most published research findings are false. PLOS Med. **2**:8, e124 (2005)
[15] Laughlin, R.B., Pines, D., Schmalian, J., Stojković, B.P., Wolynes, P.: The middle way. Proc. Natl. Acad. Sci. U.S.A. **97**:1, 32–37 (2000)
[16] Mickulecki, D.: Network thermodynamics and complexity: a transition to relational systems theory. Comput. Chem. **25**, 369–391 (2001)
[17] Nicosia, V., De Domenico, M., Latora, V.: Characteristic exponents of complex networks. Europhys. Lett. **106**:5, 58005 (2014)
[18] Nuzzo, R.: Scientific method: statistical errors. Nature News **506**:7487, 150–152 (2014)
[19] Pascual, M., Levin, S.A.: From individuals to population densities: searching for the intermediate scale of nontrivial determinism. Ecology **80**:7, 2225–2236 (1999)
[20] Tasdighian, S., Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Palumbo, P., Mei, G., Di Venere, A., Giuliani, A.: Modules identification in protein structures: the topological and geometrical solutions. J. Chem. Inf. Model. **54**:1, 159–168 (2013)
[21] Tellegen, B.: A general network theorem, with applications. Philips Res. Rep. **7**, 259–269 (1952)
[22] Todde, V., Giuliani, A.: Big Data. A briefing. Ann. Ist. Super. Sanità **54**:3, 174–175 (2018)
[23] Transtrum, M.K., Machta, B.B., Brown, K.S., Daniels, B.C., Myers, C.R., Sethna, J.P.: Perspective: Sloppiness and emergent theories in physics, biology, and beyond. J. Chem. Phys. **143**:1, 07B201 (2015)
[24] Turing, A.M.: Biological sequences and the exact string-matching problem. In "Introduction to Computational Biology", Springer (2006)

[25] Voosen, P.: Amid a sea of false findings, the NIH tries reform. The Chronicle of Higher education, March 16th, 2015
[26] Weaver, W.: Science and complexity. American Scientist **36**:4, 536–549 (1948)
[27] Young, S.S., Karr, A.: Deming, data and observational studies. A process out of control and needing fixing. Significance **8**:3, 116–120 (2011)

Received: 9 february 2019/Accepted: 31 march 2019/Published online: 30 August 2019

Environment and Health Department
Istituto Superiore di Sanità
Rome (Italy).
alessandro.giuliani@iss.it

Open Access. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.