CENTER FOR
COMPLEXITY
& BIOSYSTEMS
University of Milan

COMPLEXITYBIOSYSTEMS.IT
COMPLEXITY@UNIMI.IT

ISSUE #5 / OCTOBER, 2017
NEWSLETTER

SEBASTIANO VIGNA
CC&B founding member

Number and nodes: complexity in computer science

Complexity is sometimes an abused word, but certainly the Center for Complexity and Biosystems (CC&B) has been named aptly. The entities studied by our multidisciplinary group made of biologists, physicists and computer scientists have indeed in common that their behaviour arises from a number of articulated interconnections between smaller actors: knowing the behaviour of a single actor or entity, even in detail, even from a mathematical viewpoint, does not provide the right answers about the (complex) behaviour of the whole system. Caterina La Porta and Stefano Zapperi have described splendid examples in their disciplines--biology and physics. I will take the opportunity to discuss two examples from computer science.

Consider the generation of pseudorandom numbers, a quite fascinating and close-to-impossible area. We want to generate, in a deterministic way and using a small amount of finite state (few hundred bits), sequences of numbers that an observer cannot distinguish from numbers extracted from a random uniform source (generating truly random numbers, e.g., from radioactive decay, is prohibitively slow and expensive). Of course, we need to limit the power of the observer: since the system is finite, at some point necessarily the sequence will repeat, and an observer with infinite time and memory will detect that. But once the power of the observer is limited, we enter an interesting area. The surprising fact is that there are incredibly simple state transformations (a dozen operation or so) which, when repeatedly applied to the state, go through an unpredictable orbit. In this case, we start by the simplest component, but complexity arises because of iteration.

As a second example, anybody working on complex networks is familiar with the idea of centrality: the idea that in an interconnected network some nodes are more important than others. The idea of, simply, counting the number of connections of a node to express its importance is quite intuitive. In fact, if we interpret a directed connection between two nodes as an endorsement, or a vote, it can be traced back at least to the ancient Greeks, as it is equivalent to a democratic election. But as soon as we think again, we realize that this is a very crude way of evaluating the importance of a node: its importance should depend on the entire network structure. There are many ways to force this property: for example, by considering how far the node is from all other nodes in the network; or by defining recursively the importance of a node using the importance of the nodes it is connected to. These ideas, developed originally mainly by sociologists and psychologists with applications to small groups of people, have subsequently influenced bibliometry and, since the appearance of the web, information retrieval and computer science.

The sheer change of scale of the networks considered (from few dozens nodes to billions) has brought complex behaviour to the forefront. And now the same ideas are being used to characterize important neurons in the brain, or the structure of protein-protein networks. Showing, once and for all, that a multidisciplinary approach and contamination between fields is essential for understanding complexity.

The genetic signature that links obesity, diabetes and breast cancer

A novel approach to big data analysis allowed a group of researchers from the Center for Complexity and Biosystems (CC&B) of the University of Milan to identify a genetic signature shared by obesity, breast cancer and diabetes. Obesity is increasing worldwide at impressive rate and is overtaking smoking as the leading cause of premature death. In fact, obesity contributes to more than 70% of diabetes cases and it has been seen associated to some types of tumours, such as breast cancer. The link between obesity, diabetes and breast cancer is based on clinical and epidemiological evidence, but a strong confirmation from gene expression data is still lacking. This is mainly due to the high variability between patients and the limits of in vitro models, but also to the massive amount of noise that is unavoidable in any available data set, which makes difficult to reveal a clear signature from a large set of genes. "The huge amount of experiments in the biomedical field allowed to establish public databases that gather a large quantity of biological data", says Caterina La Porta – member of the CC&B and professor of General Pathology at the Department of Environmental Sciences and Policy of the University of Milan – who coordinated the research, just published on NPJ *Systems Biology and Applications*. "Merging data sets from different studies would be extremely useful to extract relevant information, but it is difficult because of what we call the batch effects", explains La Porta. "Each experiment introduces a bias in the data that is due to technical processing but is unrelated to biological factors. This means that this noise can mask any biological differences when comparing samples coming from distinct batches, and this is a problem". A problem that the researchers coordinated by La Porta alleviated with combining two techniques called singular value decomposition and pathway deregulation analysis. By doing so, they managed to identify 38 genes that are differentially expressed in adipocytes coming from obese and lean subjects. These genes are mainly linked to inflammation and immunity and well-known complications of obesity such as type 2 diabetes or fertility. Moreover, by comparing data from breast cancer tissue with healthy breast tissue, they were find to be similarly deregulated in breast cancer and obesity, confirming the strong association between the two. Some of them might thus represent interesting biomarkers for further studies on these topics, or even for prognostic purposes. "The strength of our work comes from the use of appropriate filtering and noise reduction methods that allow to mitigate batch effects. This general strategy can be naturally extended to other pathological conditions, providing a clear avenue to analyse the massive amount of data accumulating in the biomedical literature", concludes La Porta. "In this case, our approach allowed us to detect a list of genes characteristic of obesity, which are also associated to type 2 diabetes and breast cancer, with a degree of precision similar to that used to identify the Higgs Boson".

Integrative analysis of pathway deregulation in obesity
Font-Clos F., Zapperi S., La Porta C. A.
NPJ Systems Biology and Applications. 3, 18 (2017)

Three questions to...
Francesc Font-Clos

Researcher @ISI Foundation

What is your field of research?
I work on modelling and data analysis of biological systems. By training, I am a mathematician, but during my PhD I drifted into physics, biology and data science. Now I care about data integration methodologies, statistical robustness of different modelling approaches, dimensionality-reduction algorithms, etc.

The huge amount of data in biomedical literature might provide precious information for the development of future therapies. However, analysing these data and identifying actual patterns is still a challenging task. What are the main difficulties in this field of research?
I guess this brings us to the question of how to measure data. When we say there are thousands of gigabytes of biomedical data available online, we are giving a very crude, summarised piece of information. In a sense, it's like knowing the weight of your car in Kg, which doesn't tell you anything about its colour or its maximum speed. Datasets are a bit like cars: they have all this different characteristics, and you need to describe them properly to know if you can use them for your purposes.
So one of the first challenges in the field is that of identifying and acquiring the datasets you want to use. Both IT skills and biological understanding of the problem are equally important, so data acquisition is never a one-man job – you really need an interdisciplinary team. Then comes the problem of data integration and batch effects removal, in which we have worked a lot and I think given some substantial contributions. Basically, when you merge different datasets a special kind of noise shows up, and it gets mixed in with the biological patterns of interest. We use mathematical techniques to discern noise from signal and it is a bit like removing the background noise from old recordings: you want to get rid of the crackling noise without damaging the melody!

What are the main possible outcomes of your research and what impact could they have on biological and medical research?
We have recently published an article in which using the techniques I was mentioning we could identify specific pathways that are deregulated in obese patients. Bear in mind that none of the datasets we used could reveal those results on its own, so it was really the merging process that allowed the hidden patterns to emerge. If we are able to generalize our method to other cases, and that is something we are working in, then I think there are good chances of finding more of these hidden results.

Interview with...
Leonid Schneider
Independent science journalist

Science's self-administered poison.
This is how Leonid Schneider describes scientific misconduct. Schneider is an independent science journalist and he knows well the world of research: he worked for almost thirteen years in biomedical research on molecular cell biology, stem cells and cancer. He knows about laboratory protocols, techniques, and dynamics. He knows how to read a scientific paper and how to spot inconsistencies and contradictions. Or even something worse.
He has a blog, For better science, where he publishes his investigations on research integrity and academic publishing in life sciences and biomedicine. And in September, he was invited by the Center for Complexity and Biosystems to give a talk about research misconduct at the University of Milan.
In his investigative career, Schneider collected different examples of misconduct, from plant biology to epigenetics. In Milan, he told the case of Paolo Macchiarini, thorax surgeon, stem cell pioneer, former clinic head at the Careggi University Hospital in Florence and professor at the Karolinska Institute. Since 2008, he transplanted patients damaged by injury, cancer or other disorders with cadaveric and plastic tracheas. In both cases, tracheas were seeded with bone marrow cells, which Macchiarini said would have helped the transplants to act like biological tissue.
However, discrepancies between the medical records of Macchiarini's patients and the results he published on a paper on The Lancet were found. Moreover, all of his patients, into whom he transplanted artificial tracheas, without having first tried the method in animal models, are either dead or in permanent care.
"Suppressed evidence turned up that the Italian surgeon committed clinical and scientific misconduct as well as other acts which may amount to fraud or actual crime", explains Schneider.
But this is not the end of the story. "Macchiarini was almost immediately publicly rehabilitated by the Karolinska Institute directorate", continues Schneider. "It is thanks to a documentary from Swedish journalists that the scandal got an international media attention, as well as the most unlikely source, the glamour magazine like Vanity Fair".
This media storm was followed by an external investigation concerning the Macchiarini case, which led to some significant resignations and in the surgeon himself losing his funds and his laboratory at the Institute, in what was described as "Karolinska's Ethics Chernobyl". It turned out that Macchiarini ignored all ethical standards when he performed these operations and the material used for synthetic tracheas has never been tested in humans.
Such a scandal is not the only example of scientific misconduct and raises a key question. "How did it get that bad? Who is it to blame, Macchiarini as a crooked character or the whole system that allowed him to perform his flawed research?", states Schneider. "The answer is both. The surgeon took advantage of the hype and greed in biomedical research. He managed to charm different institutions with promises of fame and money, and left them when things got too hot. And these institutions refused to delve into proper investigations until the scandal was too big. Even then they were mostly covering up".
And then there is the problem of scientists manipulating data, which may take different forms, from the occasional cherry-picking or use of small sample size to the infamous p-hacking (statistics trickery to obtain significance) or the selective deletion of entire sets of outliers results.
"Scientists occasionally help data to fit their theoretical model", says Schneider. "The most common way to do that is through selective data acquisition or the omission of critical controls, which is a biased approach to research. And then there is the deliberate falsification of data, for instance by forging figures and graphs with Photoshop, which is a fraud." Luckily, the former is very rare, but not so rare as people imagine.
"There are two main reasons for doing that", explains Schneider. "One is to prove a preconceived theory against lack of experimental evidence, which in turns leads to irreproducible findings, pollution of scientific literature, and suffocation of correct theories. The other is to scoop a competitor, or his or her unpublished discovery. Which in turns leads to dishonestly acquired fame, funding accumulation and the domination of a research field. The findings are however reproducible then, which is good for the fraudster".
The problem is that research institutions are not always quick and prone to move against these cases of misconduct. Schneider denounces a climate of fear and a coalition of silence, due to the fact that science is at the same time a cooperative and competitive effort. It may thus happen that a researcher has to cooperate even with fraudsters, in order to find funding to keep his or her laboratory alive. And the hunt for funds may become extremely ruthless, especially in fields like biomedical research, where big money and great expectations are. Larger funding means greater competition, and funds are necessary to perform high-level research. Which is necessary to publish papers on top journals. Which in turns are necessary to have a CV strong enough to get more funding. It is a kind of vicious cycle from which it is difficult to get out. Moreover, false misconduct allegations are sometimes used to secretly damage competitors, critics and rogue ex-employees.
"In such a context, the key to restore scientific and research dignity is transparency", concludes Schneider. "Transparency means, for research institutions, to be quicker to investigate and, if necessary, to act against demonstrated cases of misconduct. Forged and manipulated papers need to be retracted quickly and manifestly, and the culprits must be pushed away from research".

Training a computer to assess sperm quality

Training a machine to classify sperms based on their physical traits: this is the task accomplished by a group of researchers from the Center for Complexity and Biosystems of the University of Milan, who just published the study in ***Scientific Reports***.
A task that might be of great help in the field of reproductive medicine. The presence of abnormalities, such as a large or misshapen head or a crooked or double tail, might affect the ability of the sperm to reach and penetrate an egg. For this reason,

sperm morphology is one of the factors that are examined as part of a semen analysis to evaluate male infertility. And such evaluation is usually performed by experts with a trained eye, who look at the semen under a microscope and classify sperms based on their aspect. However, due to the increasing amount of available digital images, it is becoming important to develop automatic techniques of classification and diagnosis. To do such a thing, it is then necessary to develop reliable automated methods for cell morphology assessment. Objective tools only exist for sperm motility assessment, but not for sperm morphology, so that subjective morphology sperm cell assessment is still the standard in laboratories. "Machine learning-based intelligent systems could play a pivotal role to reach this goal", says the biologist Caterina La Porta, from the Department of Environmental Sciences and Policy, who coordinated the research. "These systems can train themselves to learn the patterns of the data we provide them and produce a prediction model. The final goal is then to be able to automatically classify a data set with unknown labels".

The researchers focused their attention on the morphology of the acrosome, an organelle with the shape of a head-cap that covers the sperm nucleus. They used a large amount of images of mouse sperms to perform a three-dimensional digital reconstruction of their acrosomes, and then compute a series of parameters such as volume, surface and local curvatures. Finally, they analysed these traits by machine learning and compared them with the ground truth provided by a direct assessment by eye. The algorithm spotted differences that an expert eye was not able to distinguish, and its classifications were corrected in 73% of trials – a higher percentage than those obtained with other methods. "We have proposed a general strategy to classify acrosomes during the course of sperms development, according to their morphological features", concludes La Porta. "This could help solve the relevant clinical issue of quantifying the percentage of sperm cells with normal acrosome and therefore assess fertility".

Probing spermiogenesis: a digital strategy for mouse acrosome classification
Taloni A., Font-Clos F., Guidetti L., Milan S., Ascagni M., Vasco C., Pasini M. E., Gioria M.R., Ciusani E., Zapperi S., La Porta C.A.
Scientific Reports. 7, 3748 (2017)

The statistics of fluctuations in amorphous materials

There may be a general physical law behind the way many materials change irreversibly their own shape when pressed or pulled. This is what a group of scientists from the Center for Complexity and Biosystem (CC&B) of the University of Milan may have discovered while studying the deformation processes in amorphous solid materials.

The structure of most solids is crystalline, which means that their atoms or molecules are organised in a regular and periodic manner. On the other hand, liquids have a completely random structure. There is also an intermediate case, represented by amorphous solids: they lack the crystalline structure of many solids but, differently from liquids, their randomly arranged atoms cannot flow easily. The most famous examples of amorphous solid is glass, but the categories also includes gels, thin films and many polymers, some of which may consist of both crystalline and amorphous regions.

When an object is exposed to external forces, something happens within its atomic structure: a series of micro-events that eventually results in the deformation, or even the wreckage, of the object itself. Material scientists know this, and they also know that different materials deform in different ways.

"There are two kinds of deformation, plastic and elastic", explains Stefano Zapperi, professor of theoretical physics at the University of Milan, head of the CC&B and coordinator of the research, just published on *Nature Communication*. "A deformation is plastic when the shape of a material changes in an irreversible way but without breaking. Conversely, when a material can revert to the initial state, its deformation is called elastic. For this work, we chose to focus on plasticity in disordered solids".

In 2011, Zapperi obtained an Advanced Grant from the European Research Council to conduct research on these topics, with a project called SIZEFFECTS. Within this project, he and his colleagues developed a computer model that mimics the deformation of an amorphous material following the exposure to different kinds of external forces, like tension and bending, with different intensities and directions.

"Our aim was to compare the material response at the microscopic scale to its deformation behaviour at bigger scale", says Zapperi, "and to analyse this relationship from a statistical point of view. Our findings provide compelling evidence for generic dynamics of the statistics of fluctuations. Clearly, there is some sort of universality behind all these plastic deformation processes in amorphous materials, differently from previous explanations", concludes Zapperi. "It is not different from what happens in earthquakes, where the sum of small and large events eventually results in the motion of a geological fault. We believe that our results might be of great help for future applications like materials engineering and design, especially for the production of metallic glasses".

Paper references

Universal features of amorphous plasticity
Nature Communications
Zoe Budrikis, David Fernandez Castellanos, Stefan Sandfeld, Michael Zaiser & Stefano Zapperi
DOI: 10.1038/NCOMMS15928

>>> UPCOMING EVENTS

Paolo Vineis
Imperial College, London
Cancer: "bad luck" or environment? The contribution of exposome research

November 24st 2017
14.00 — BS Room
via Celoria 26, Milano

Diego Liberati
CNR
Quantitative systems bio-physiology for precision personalized medicine: from data mining through signal processing to mathematical modeling

November 30st 2017
12.30 – TBA

CC&B is a Coordinated Research Center at the University of Milan
Research within CC&B is supported by the European Research Council
CC&B cooperates with the ISI Foundation www.isi.it