Math. Model. Nat. Phenom.
Vol. 10, No. 3, 2015, pp. 186–205
DOI: 10.1051/mmnp/201510314

Overcoming Complexity of Biological Systems: from Data Analysis to Mathematical Modeling

A. Zinovyev[1,2,3] *
[1] Institut Curie, 26 rue d'Ulm, Paris, 75248 FR
[2] INSERM U900, Paris, FR
[3] Mines Paris Tech, Fontainebleau, FR

Abstract. The problem of dealing with complexity arises when we fail to achieve a desired behavior of biological systems (for example, in cancer treatment). In this review I formulate the problem of tackling biological complexity at the level of large-dimensional datasets and complex mathematical models of reaction networks. I show that in many cases the complexity can be reduced by using approximation by simpler objects (for example, using principal graphs for data dimension reduction, and using dominant systems for reducing complex models). Examples of dealing with complexity from various fields of molecular systems biology are used, in particular, from the analysis of cancer transcriptomes, mathematical modeling of protein synthesis and of cell fate decisions between death and life.

To Alexander Gorban, my scientific supervisor, on his 60th birthday

Keywords and phrases: complexity, biological, systems, modeling, data analysis, cancer

Mathematics Subject Classification: 92-00, 92-08, 92B05, 92C42, 62P10

1. Introduction

We struggle with the complexity of biological systems, which are resistent to our attempts to understand and, hence, efficiently manipulate them.

The notion of complexity is multi-faceted since many different notions hide behind the "complexity" term: the number of elements or connections between them, number of intrinsic degrees of freedom, non-triviality of behavior, non-linearity of mathematical equations, difficulties with abstraction, etc. Some researchers associate the notion of complexity with non-linearity or large dimension. Others connect complexity with emergence and self-organisation [54]. Some point out that the main challenge on this way is to distinguish complicated and truly complex systems, though no consensus view on the nature of this distinction exists in the community. Complexity of biological systems is tightly connected to their robustness and the history of their evolution [1, 67].

In his "millennium" interview, Stephen Hawking said *"I think the next century will be the century of complexity"*. What is meant here is that in the XXI century most of scientific effort will be devoted

*Corresponding author. E-mail: andrei.zinovyev@curie.fr

not to discovery of new scientific laws; but rather, knowing the fundamental laws, to understanding how complex systems are assembled and function. Citing another paper on complexity theory, *"a new scheme of actions became dominant in the struggle with complexity. The complexity is recognized as the gap between the laws and the phenomena. We assume that the laws are true. We can imagine a "detailed" model for a phenomenon but because of complexity, we cannot work with this detailed model. We can imagine a detailed kinetic equation for a reaction network but cannot find reaction rate constants and cannot work with this large system even if it is true"* [30].

In this view, complexity is presented as an obstacle for the human mind, equipped with modern technology, to interpret the behavior of complex systems based on a set of simple laws with deductive reasoning, whether it uses common sense or computational approaches.

From an abstract point of view, there are two methods for dealing with complexity: on the one hand those based on model or dimension reduction and on the other those based on self-averaging [1, 22]. To understand this dichotomy, we can think of a complex phenomenon as an object existing in a multi-dimensional space (e.g., a set of points or trajectories or a vector field). Our perception of this object is inevitably low-dimensional because our mind is organized by representation of our motion in three-dimensional space and the convenient static visualisation is two-dimensional. Therefore, we can represent our perception as a projection of the object from high-dimensional to low-dimensional space. A biological function can be also considered as a projection of its high-dimensional microscopic detailed description onto a low-dimensional space where it is manifested at macroscopic level.

The reducible complexity model states that despite the fact that the complex object is embedded in a high-dimensional space, intrinsically, it remains low-dimensional with a relatively small number of degrees of freedom (case of injection of a low-dimensional object into multidimensional space). Reducible complexity of data often means existence of *lower-dimensional principal manifolds*. Reducible complexity of dynamical systems is manifested in low-dimensional intrinsic structure of their attracting sets or existence of *low-dimensional invariant manifolds* [23, 24]. Another frequent type of reducible complexity is a system's structure following some relatively simple organisational principle. One of the most common principles is the hierarchical organisation. In biological networks, this type of hierarchical reducible complexity is revealed in the existence of modules, compartmentalisation, multiple concentration and time scales [61]. In physiology, it can be seen as the construction of an organism from organs, tissues and cells.

By contrast, self-averaging complexity is associated with truly high-dimensional objects that do not possess any intrinsic low-dimensional simple structure. However, projections of this object on most of the low-dimensional screens will look very similar. A good mathematical metaphor for this type of complexity is a multi-dimensional shape (a hypercube, a hypersphere) uniformly sampled by points. After projection on any two-dimensional plane, most of the points will be located very close to the center of the projection distribution. Moreover, the distribution of the projected points will be very close to a normal one (i.e. Gaussian). In statistical physics, this corresponds to the well-known Maxwellian distribution. Generally speaking, self-averaging is a manifestation of the Gromov's measure of concentration phenomenon: truly high-dimensional objects look very small (concentrated) after projection onto a low-dimensional space and most of the distributions become almost normal after projections on the low-dimensional subspaces [39].

Wild complexity cannot be simplified neither by reduction nor by averaging. There is no good lowdimensional screen to observe objects possessing this property: different projections will give different representations of the object, with new features. One cannot dissect this complexity with levels because there is no clear separation between them, they all coexist and penetrate into each other. There is no time or space scale separation; most of the processes are happening at the same time and everywhere with strong between-scale coupling. In wild complex systems, many local perturbations produce global effects which might be very different from one perturbation to another similar perturbation.

Examples of wild complexity are actually rare, because its definition comes from a negation: it is a complexity that cannot be reduced or averaged whereas any simple illustrative example will be already

reduced or averaged. Probably, one of the few examples can be found in the collective neuron excitation dynamics of our brain. [45] developed a computer brain model with one million multi-compartmental spiking neurons and a billion synapses. The model is calibrated to reproduce known types of responses recorded *in vitro* in rats. Computer simulations of this model show overwhelmingly complex dynamics characterised by global excitation-like responses, spontaneous activity, sensitivity to changes in individual neurons, functional connectivity on different scales. The complexity of this model can be tentatively characterised as wild.

A superficially similar, but actually distinct notion is that of irreducible complexity. As used in evolutionary theory, this notion is historically associated with intelligent design ideas (in particular, by Michael Behe). The irreducibly complex systems were defined as *composed of several well-matched, interacting parts that contribute to the basic function, wherein the removal of any one of the parts causes the system to effectively cease functioning,* i.e. as extremely fragile systems. The notions of wild and irreducible complexities are related to different problem statements: the first is related to impossibility to find a simple representation while the second is about impossibility to "simplify" by removing a part.

In the field of systems biology, complex datasets (e.g., genome-wide measurements, omics data) and complex models (e.g., global interaction networks) are appearing at an increasing rate. An important question is to what extent we are able to simplify them in order to understand and control the biological system's behavior. Below I present several projects in which this question was addressed in different ways, and specific computational or mathematical tools that have been used to answer it.

2. The Big Data of Molecular Biology

2.1. Curse of data quantity and of data dimensionality

The current state of molecular biology and genetics is characterized by unprecedented influx of quantitative data. Twenty years ago, bioinformaticians used to deal only with relatively small and fragmented collections of genetic sequences, to which they applied careful and intellectually challenging analyses (such as estimation of the mutation rates). Modern biotechnology allows generating data at an exponential rate, the exponent of which is larger than that in the Moore's law, which describes the growth of the computational power in hardware [1]. This growth rate inevitably leads to the situation when storing and analyzing the data becomes more expensive than producing them.

The technological challenge of efficient storage, compressing and pre-treating the data is accompanied by the scientific challenge of using the data in order to extract useful knowledge from them. A biological sample can be characterized by increasing number of quantitative features. In the beginning of the microarray era one had few tens of thousands of measurements per sample. Using modern sequencing techniques, the number of extractable numerical features such as the number of RNA counts, different forms of RNA, mutations or epigenetic modifications of DNA, variations in DNA sequence, grew by several orders of magnitude. For example, a sample from tumour biopsy can be characterized by the corrupted genome sequence and abberant DNA structure in tumoral cells and averaged concentrations of a hundred of thousands of biologically relevant molecules in cells and their microenvironment. This leads to a methodological problem known as "small n, big p" problem (where n is the number of samples and p is the number of quantitative features describing the sample).

Imagine that a biological sample is a point in a multidimensional space whose dimension equals to the number of measured numerical features. A collection of samples is represented as a cloud (discrete distribution) of points in this space. If this distribution is truly multidimensional (there is no low-dimensional subspace around which this cloud is concentrated) then the average distance between a point and its closest neighbor and the average distance between all pairs of points are comparable. Imagine a ranking of points from a selected point sorted by distance. In the truly multidimensional situation, adding a small amount of noise to the position of data points risk to change this ranking drastically. This is the essence of the curse of dimensionality. Most of the statistical methods, including the simplest k-nearest neighbors classifiers and linear predictors, performs very poorly in such conditions, their parameters are

too sensitive to imprecisions in data measurements or to random removal of a small fraction of samples from the dataset.

One of the manifestations of the curse of dimensionality is difficulty of finding recurrent patterns in the data such as frequently repeated events. In cancer, for example, most of the mutations which can in principle serve as biomarkers of the treatment success, are present at very small frequencies. This hampers evaluating their statistical significance and requires an enormous (and not always feasible) amount of biological samples.

Typical methods which were developed to overcome this type of complexity are feature selection methods, regularization and dimension reduction. I will consider in more detail below a particular family of methods of dimension reduction, connected to construction of principal manifolds.

2.2. Linear and non-linear data dimension reduction

For data approximation, the notion of the mean point can be generalized by more complex types of objects. In 1901 Pearson proposed to approximate multivariate distributions by lines and planes [59]. This idea gave born to the Principal Component Analysis (PCA) which nowadays is a basic statistical tool. Principal lines and planes go through the "middle" of multivariate data distribution and correspond to the first few modes of the multivariate Gaussian distribution approximating the data. See application of the method of principal components to visualize the structure of cancer genome in Figure 1.

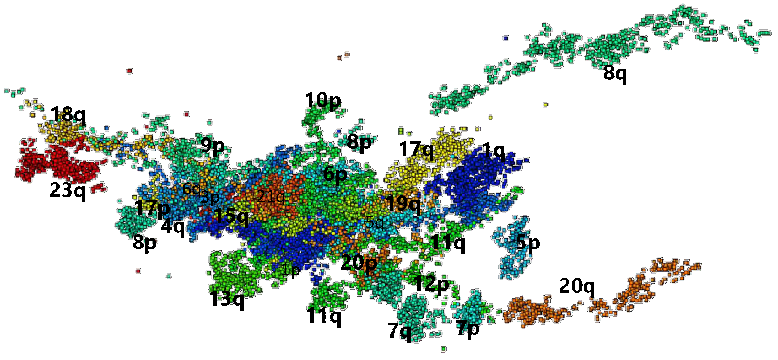

FIGURE 1. Visualization of cancer genome. This is a PCA plot of genome locuses characterized by their profiles of copy number changes in a series of 160 breast and ovarian cancer cell lines, available in Broad Cancer Cell Line Encyclopedia (http://www.broadinstitute.org/ccle/home). Different colors designate different chromosomes (starting from blue, 1st chromosome, to red, 23rd, or X-chromosome, p and q signify two chromosome arms). The horizontal PCA axis is approximately associated with the frequency of gains (on the right) and losses (on the left). The vertical PCA axis does not have evident interpretation. Those chromosome arms located closely on the plot have similar profiles of gains and losses. The data for this analysis were kindly provided by Dr. Tatiana Popova.

Data approximation methods can be extended for other types of principal objects: first of all, to non-linear manifolds of various topologies. Using this principal object type inspired the creation of the method of elastic maps for constructing principal manifolds [25,31,32,36,69]. The method uses an analogy between smooth surface and elastic membrane and consists in optimizing a penalized data approximation functional which can be interpreted as elastic energy of a system of springs. The method is implemented in several software packages available online (http://bioinfo-out.curie.fr/projects/vidaexpert, http://bioinfo-out.curie.fr/projects/elmap,
http://bioinfo-out.curie.fr/projects/vimida).

The method of elastic maps was applied in bioinformatics to many different tasks. It was used to visualize the universal 7-cluster structure of bacterial genomes [37, 38] and the structure of codon usage in genomes of various organisms [12, 68]. Elastic maps allow approximation of molecular surfaces of complex molecules and visualizing them [32]. It is routinely used for analysis of microarray data in cancer biology [34, 35, 68] and in biology of microorganisms [14]. The method of elastic maps was applied in quantitative biology for reconstructing the curved surface of a tree leaf from a stack of light miscroscopy images [19].

In [25] we developed a series of systematic tests to quantify the quality of data projection onto a non-linear principal manifold in terms of preserving the structure of small and big distances and conserving class relations between data points. These methods were applied to transcriptomic datasets and it was shown that two-dimensional non-linear principal manifolds perform systematically better than two- or three-dimensional linear principal manifolds, which is an indication of the intrinsic non-linear structure of transcriptomic data.

Recently, the idea of non-linear principal manifolds was revived in molecular biology data analysis: for example, for describing the progression of colon cancer from hyperplasia to metastasis [18]. Under different names (e.g., Wanderlust algorithm), the idea of principal curves was applied for approximating the large-scale data in development studies, including applications to single-cell molecular profiling [2].

Further extension of the method led to the idea of approximating datasets by arbitrary graphs (principal graphs) [28,29,34,35]. It was suggested to define such graphs by applying topological grammars [28]. The simplest possible grammar leads to the method of *principal trees*. A principal tree allows approximating complex datasets having intrinsic branching structure (such as molecular profiles of cell lineage data), see Figure 2.

Recently, we used approximation of datasets by principal trees in order to evaluate the complexity of the datasets [73]. We introduced three natural types of data complexity: 1) geometric (deviation of the data's approximator from some "idealized" configuration, such as deviation from harmonicity); 2) structural (how many elements of a principal graph are needed to approximate the data), and 3) construction complexity (how many applications of elementary graph transformations are needed to construct the principal object starting from the simplest one). We computed these measures for several simulated and real-life data distributions and showed them using the "accuracy-complexity" plots, helping to optimize the accuracy/complexity ratio.

Reduction of the data dimension represents the first step towards controlling biological complexity. Reducing the number of variables by eliminating uninformative ones or lumping the variables into weighted combinations helps in extracting the knowledge from the data for hypothesizing and further analyses.

2.3. Blind source separation of biological signals

One can be interested in studying the numerical object features (such as expression of a particular gene) across a number of biological samples. This situation can be represented as a distribution of p points in n-dimensional space, where $p \gg n$. Such representation helps answering the following question: *which biological or technical factors affect the genome-wide measurements* (for example, the transcriptome)? The biological factors can be activities of transcription factors or other various influences coming from a particular intercellular context or from the environment. The technical factors can be various biases connected to the preparation or even extraction of the samples, or to the technology used. The combi-

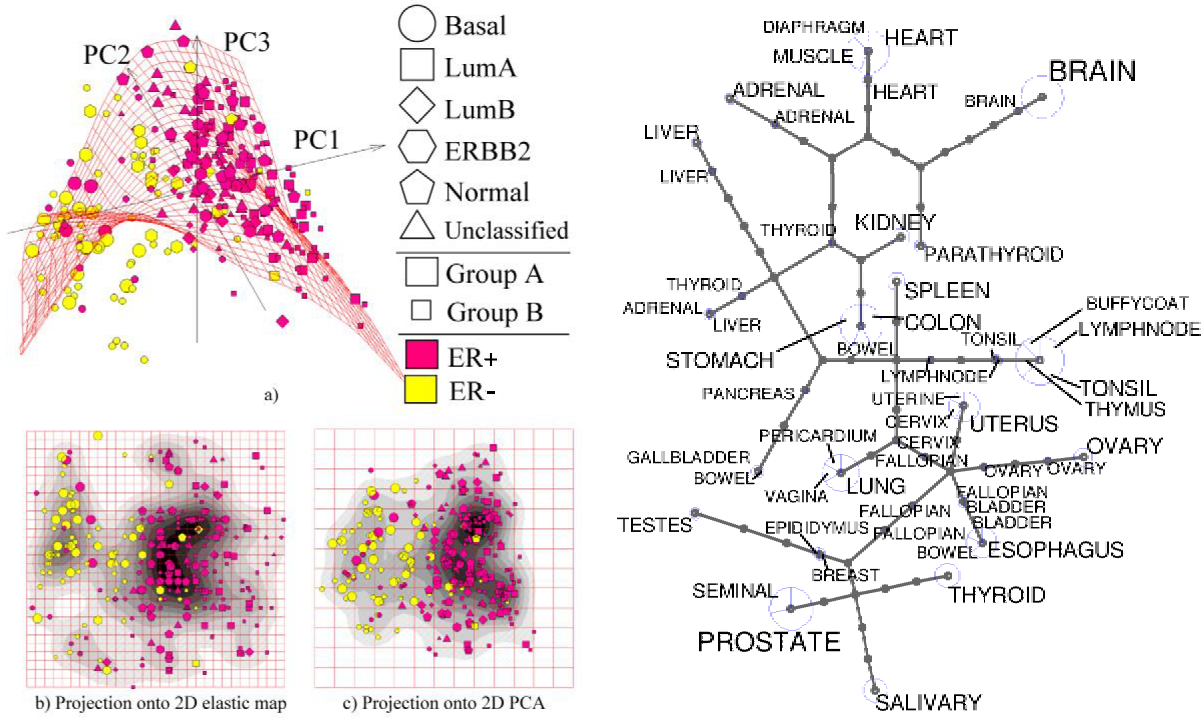

FIGURE 2. Linear, non-linear and graph-based approximations and visualizations of biological datasets (microarray data, transcriptomes). Left: comparison of linear and non-linear principal manifolds for visualization of the structure of distances between breast tumor samples (from [33]). Right: visualization of a dataset of normal tissue samples (from about 100 tissues), the structure of the distances is represented by a principal tree with 2D layout called "metro map" (from [25])

nation of factor activities regulates gene expression in a complex (and unknown) function. As the first approximation, we can assume that this function is linear:

$$Expression(gene_i, sample_s) \approx \sum_{j=1..m} a_{F_j}^{gene_i} Activity_{F_j}(sample_s). \tag{2.1}$$

Linear decomposition of the gene expression table depends on conditions on gene responsiveness coefficients $a_{F_j}^{gene_i}$ connected to a particular signal (factor) F_j and the activities $Activity_{F_j}$ of this signal in a sample s. A condition of orthogonality or uncorrelatedness leads to the standard Singular Value Decomposition problem. A requirement of statistical independence of the signals leads to the Independent Component Analysis (ICA) method, which was initially developed for the signal processing field (see a review of the application of this method for cancer research in [71]).

We applied this approach to comprehensively characterize the activity of biological factors in the transcriptome of bladder cancer, and compared these factors to the signals that can be detected in other cancer types (Figure 3).

We interpreted the inferred independent components by studying their association with predefined gene sets, and also with the clinical, pathological and experimental data, mutations, and copy number alterations available for the these tumors. We also applied ICA to publicly available gene expression datasets, composed of six series of bladder cancers and 13 series of cancers of other types (in total, 6671 tumour samples of 9 different cancer types were analysed). Our results showed that ICA allows: the

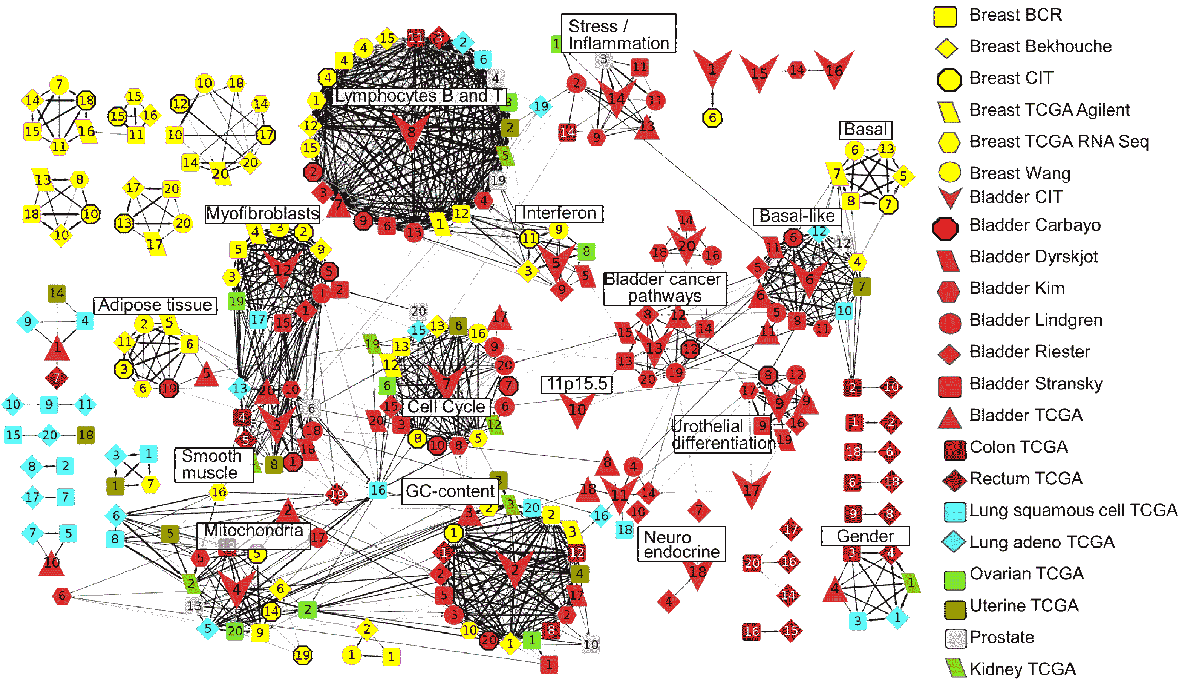

FIGURE 3. Comprehensive characterization of transcriptomes of collections of tumour samples in different solid cancers (from [4]). The image is a correlation graph representing the correlations between independent factors identified in various datasets (node shapes) and cancer types (node colors). The text annotations denote several clusters of components (cliques) representing reproducible biological signals which origin is relatively well-understood. For example, LB&T represents a set of components connected to the reaction of immune system (presence of lymphocytes, B- and T-cells in the tumour sample).

identification of gene expression data which are influenced by technical biases; the identification of processes that were either cancer type-specific or common to several types of cancer, the characterization and the comparison of predefined tumor subgroups and the identification of genes involved in carcinogenesis. One of our results involved the prediction of a potential pro-tumorigenic role of the nuclear receptor PPARG (peroxisome proliferator-activated receptor gamma) in bladder tumorigenesis. This prediction was confirmed by functional studies [4].

Meta-analysis of cancer transcriptomes with use of blind source separation methods allows us to answer important questions, such as what are the most reproducible factors affecting cancer transcriptome, among various types of cancers? For example, it can be demonstrated that the factor associated with organization of tumor microenvironment (extracellular matrix properties, expression of metalloproteases, various aspects of cell-cell and cell-matrix adhesions) is the most reproducible among breast cancer transcriptomic datasets. At the same time, in a set of four solid carcinomas (breast, ovary, lung and prostate cancer), the most reproducible common signal is associated with immune response and infiltration of T-cell in the tumoral tissue.

In all studies dealing with computing principal manifolds for transcriptomic datasets, it was demonstrated that there are relatively few (about twenty) factors shaping expression profiles of tumoural samples, i.e. the typical intrinsic dimension of the manifold where the tumour transcriptomes are located is small. The embedding of gene copy number profiles of cancer genomes has an effective dimension around

15 (Figure 1). Therefore, when studying transcriptomic or genomic profiles, we are typically dealing with a reducible case of complexity.

3. Mathematical models of biological systems: complex of complicated?

3.1. Asymptotology of reaction networks

Biological systems show complex behavior, where complexity is manifested in the difficulty to predict what will happen with a particular biological system or its element as a result of a perturbation. Biological systems are constructed from microscopic parts which are not directly observable. Hence, any observation on their behavior represents a complex function integrating many molecular components and their interactions. Therefore, it is inevitable to introduce abstracted models of reality connecting the microscopic and macroscopic behavior. The notion of a model is well established in molecular biology. Mathematical approach provides a formal language in which such models can be formulated and analyzed. The model serves as a "theoretical microscope" allowing to look mentally at the unobservable molecular mechanisms from their macroscopic manifestations assuming the validness of the basic laws that govern the molecular interactions.

The most fundamental basis for constructing the formal biological models are the laws of (bio)physics and (bio)chemistry. One can assume that these laws are well-known. However, these laws by themselves do not explain functioning of biological systems. Galileo pointed out that that the laws of physics tell you there are no flying elephants but they cannot help you to draw the path that brought walking elephants to Earth [40]. The main challenge for molecular biology is to connect the basic laws of nature to complex organization of biological organisms.

Most mathematical methods for modelling molecular mechanisms regulating complex behavior of living cells are based on formal methods of chemical kinetics developed for studying chemical or biochemical systems. Historically, these methods were introduced to deal with relatively small and well-defined systems of chemical reactions. It is clear now that the biochemical cascades involved in cellular signalling can be characterised by large (a few hundreds of components) and extra-large (more than one thousand components) complex hierarchical structure and by multiscale temporal and spatial behaviour. Modelling large biochemical networks, based on the standard mathematical approaches, faces obstacles such as incompleteness of network description (structural and parametric), lack of exact knowledge of kinetic parameters, fuzziness of borders of the classical biochemical pathways, intensive pathway cross-talk, instabilities and poor scalability of numerical solvers. These peculiarities are not explicitly taken into account in the general mathematical methods applied in most of systems biology today. This crucially limits the success of mathematical modelling in the field of systems biology of human diseases such as cancer.

Why do we need large and complex network models? Constructing large models of biochemical reactions gives serious advantages for realistic mathematical modelling of signalling pathways. Firstly, it affords the possibility of representing the complexity of molecular mechanisms without neglecting or a priori over-simplifying their components. This could be essential for understanding variability of individual response to treatment that cannot be captured by fixing a generic simple framework. Secondly, it allows the combination of separate existing mathematical models into comprehensive master models, when interaction between various biological mechanisms (for example, cell cycle and apoptosis) is known and cannot be neglected. Thirdly, it allows easier comparison between modeling and high-throughput data, since these data are available at the level of individual and elementary molecular entities. Lastly, determination of kinetic parameters can be less difficult for elementary molecular mechanisms as opposed to more abstract mechanism representations.

However, complex mathematical models can be as intractable as the biological systems themselves. With a very complex model, a researcher can observe possible system behaviour from numerical simulations, but he or she is not able to predict changes in the model dynamics as a response to changing its parameters values. Empirically, each particular numerical simulation can show very simple dynamical

properties (for example, almost linear relaxation dynamics). In this case, the dynamics can be described by very few key parameters or their combinations. The problem is in that the recipe for finding these combinations is usually not known.

One of the most striking observation in mathematical modeling of biological systems is frequent discrepancy between the complexity of data (simple relaxation curves or non-systematic measurements with wide confidence intervals) and the complexity of the mathematical models aimed to reproduce these data. As a result, the parameters of complex models fitted to experimental data can be highly undetermined [44]. The solution could be to reduce the complexity of a mathematical model to the level of complexity of experimental data, and fit only the necessary key parameters, which are usually some functions of the parameters in the detailed model. All this dictates a need in developing methods of computational modeling allowing to work with large and incompletely characterized networks of biochemical interactions.

Methods of simplifying complex chemical reaction networks and their dynamics can serve as a basis for the theory of asymptotology of chemical reaction networks [27]. Following [48], asymptotology is *"the art of describing the behavior of a specified solution (or family of solutions) of a system in a limiting case... The art of asymptotology lies partly in choosing fruitful limiting cases to examine... The scientific element in asymptotology resides in the non-arbitrariness of the asymptotic behavior and of its description, once the limiting case has been decided upon"*. Known methods of model reduction (quasisteady-state, quasiequilibrium asymptotics, lumping approaches, methods based on limiting reaction steps) are examples of finding simple asymptotic solutions of complex chemical kinetics equations, hence, they form the theoretical basis for asymptotology of reaction networks.

The interpretation of the asymptotology principles might be the following. Let us imagine a complex model of a large biochemical reaction network, characterized by a particular set of kinetic parameter values and some dynamics. Asymptotology claims that in many cases the dynamics of the model will not be as complex as it might be expected from the model size or its non-linearity. For example, in any given moment of time, most of the subsystems in the model will be characterized by the simplest quasi-linear relaxation, and only a minor subset of interactions will be described by truly non-linear equations. However, this subset might change in a different time point: a complex system "walks" through its simpler subsystems.

Asymptotic description of the dynamics consists in 1) for a given set of parameters, decomposing the complex dynamics into periods (epochs), each of which can be described in relatively simple terms (asymtptotic, based on neglecting some parameters or quantities); 2) for all admissible sets of paratemers, listing a set of possible asymptotic behaviors. Asymtotology provides tools for constructing this description for some classes of complex and large networks. Ideally, asymptotic solutions should be simple or trivial, and even analytically tractable. This allows understanding the system properties and predicting possible changes of dynamical properties of the model as a response to a change of parameter values.

In biological terms, it means that any particular system response is characterized by a relatively simple sequence of events which are tractable and comprehensible. The complexity of the biological system arises from its ability to respond to many different types of signals and perturbations, thus, holding a capability of many simple but different dynamics. Conceptually, it remains unclear if this view is applicable to the functioning of real biological reaction networks. If it is true then the biological complexity can be dissected as a superposition of relatively simple behaviors. In the opposite situation, the biological complexity is "wild", and not tractable. Imagine a well-defined large biochemical system, and a sufficient amount of experimental data to completely determine its parameters. The form of the distribution of these parameter values will give a first hint to the system complexity: if they are distributed almost uniformly on the log scale then the complexity is most probably "dissectable". Secondly, one can investigate the properties of the vector fields describing the system dynamics. If it will be characterized by existence of low-dimensional attractive slow manifolds then the complexity is "dissectable". A method for making such a test was suggested in [64].

Experience shows that most of the existing mathematical models (collected, for example, in BIOMODELS database, http://www.ebi.ac.uk/biomodels/) trained on some real experimental data are "dis-

sectable" in terms of complexity [63]. However, this can reflect only our way to fit the models to data, or to the limitations of experimental techniques. In principle, it is meaningful to ask the question if the natural selection leads to "wild" or "dissectable" complexity in biochemical reaction networks. Conceptually, this question is close to studying the evolution of network robustness [1, 62, 63].

Asymptotology or model reduction methods allow overcoming biological complexity in several ways. Firstly, they allow making the model equations simpler and more tractable (analytically or numerically). Secondly, they allow determining the key model parameters and their relation to the parameters of the complete model. Thirdly, they allow dissecting model complexity into a set of simple models, and match each biological observation to a possible asymptotic model dynamics. Fourthly, they allow predicting how to switch between different asymptotic (qualitatively different) modes of behavior. Finally, they allow to decide if a mathematical model is truly ("wildly") complex of just complicated ("dissectable").

In some particular systems, this research program can be accomplished successfully: for example, a method for reducing a very large system of equations describing a network of monomolecular reactions in the case of well-separated kinetic constants was developed [62]. Interestingly, these methods have tight connection to quite abstract mathematics such as tropical algebras and "model tropicalization" [63] or the notion of dominant system in dynamical systems [27]. Below I will focus on describing one particular application of this approach to the problem of modeling the mechanisms of miRNA action on protein translation.

3.2. Mathematical modeling of miRNA-mediated translation repression

In a series of stidues [55, 74, 75], we applied asymptotology and model reduction for dissecting complexity of a complex molecular mechanism of miRNA-mediated translation repression.

MicroRNAs can affect the protein translation using nine mechanistically different mechanisms, including repression of initiation and degradation of the transcript [55]. There is a debate in the current literature about which mechanism and in which situations has a dominant role in living cells. This debate is complicated by the observation that same experimental systems dealing with the same pairs of mRNA and miRNA can provide ambiguous evidences about which is the actual observed mechanism of translation repression. Mathematical modeling is able to suggest explanation to existing controversies in the field.

To understand the effect of miRNA on translation, one needs to construct a kinetic model of translation. Two simple models of translation were developed in [56], and we started from the detailed analysis of them.

The first model represents a simple cycle of three reactions [56, 75]: 1) free ribosomal subunit 40S binds to mRNA, 2) full ribosome is assembled at the start codon, 3) mRNA is translated and ribosomes are released from mRNA. Even this simple model suggests that depending on which step of translation is limiting, the effect of miRNA can be detectable or not detectable. Thus, if 40S binding to mRNA is a rate-limiting step, but miRNA modulates the ribosome assembling step, then the effect of miRNA will not be observed in the experiment (for realistic values of the strength of the modulation). This indeed happens when mRNA with a modified artificial cap structure is used, which makes the initiation step very inefficient (hence, rate-limiting).

A Science paper [53] wrongly concluded that the absence of miRNA effect on the steady-state rate of protein synthesis, for mRNA with a modified cap structure, proves that miRNA action requires the normal mRNA cap. Mathematical modeling, however, showed that it is not possible to distinguish between two explanations: 1) miRNA action is cap-dependent or 2) miRNA acts at the ribosome assembly step (which is not rate-limiting). As we suggested, it is possible to distinguish between two mechanisms, if together with the change in the steady-state rate of protein synthesis one would measure also the relaxation time, i.e. the time needed to arrive to the new protein synthesis steady rate [75].

The second model suggested in [56] is non-linear because it explicitly takes into account the turnover of translation initiation factors and ribosomes. It would be natural to determine which reaction in this model is rate-limiting, but this is not possible. Actually, it is known that the notion of rate-limiting reaction

step is applicable to only very simple systems, while in the non-linear case the rate limiting "place" can be not a single reaction and, moreover, can change with time. Thus, in complex and non-linear systems the notion of rate-limiting step should be replaced by the *dominant system* [27]. The dominant system in this context is such a simplified dynamical system which results from the assumption that the kinetic rates in each fork of reactions are well-separated (i.e., asymptotically infinitely different in their values). Therefore, the dominant system asymptotically approximates the dynamics of the initial system. Usually, they represent much simpler reaction networks though they are not necessary subnetworks of the initial systems (some reaction rewiring is usually necessary). Example of computing dominant systems can be found in our work on studying the non-linear model of translation suggested in [56]. For this model we performed a complete asymptotological analysis as it was depicted in the previous section. We listed all possible asymptotic system behaviors, and showed how the asymptotic solutions (obtained by integrating the the dominant system) change each other in time. As a result, a semi-analytical solution for a non-linear model of translation was obtained and analyzed for possible measurable effects of miRNA [75]. As a result, we suggested experimental designs in which several mechanisms of miRNA action can be distinguished.

This work was continued with creating a new kinetic model of coupled transcription, translation and degradation. We build such a model by lumping multiple states of translated mRNA into few dynamical variables and introducing a pool of translating ribosomes [26]. In this model, it is possible to simulate all nine known mechanisms of miRNA action (Figure 4). Moreover, it is possible to consider the situation when several mechanisms of miRNA action are present simultaneously, and to predict their measurable effect.

We dissected the complexity of our translation model by applying the asymptotology approach. We found out that the model has 6 distinct asymptotic behaviors. Some of them correspond to a single mechanism of miRNA action, but most correspond to several mechanisms simultaneously. Classification of miRNA action into 6 dynamical types is more constructive from the point of view of observable variables (such as total mRNA and protein amounts, and the polysome profiles) than considering the individual molecular mechanisms. This classification shed light onto the controversies existing in interpreting the results of experiments. Indeed, as our analysis showed, the same mechanism of miRNA action can lead to different translation dynamics depending on the concrete distribution of parameter values and to confusion in the interpretation.

Based on our analysis, we suggested precise recipes (kinetic signatures) on how to distinguish between different mechanisms of miRNA action from experiments on observing the dynamics of the amounts of mRNA, protein and the average number of ribosomes translating one mRNA [55]. These theoretical results await experimental validation. In conclusion, studying the mathematical models of complex molecular mechanism of translation showed that the asymptotology approach is an efficient tool for dissecting the complexity into relatively simple scenario, each of which can be analyzed and understood unlike the comprehensive model: in other words, overcoming complexity of mathematical models is possible.

4. Cancer as a complex system

4.1. Google Maps of cancer biology: Atlas of Cancer Signaling Network project

Most biological knowledge exists only in narrative form, dispersed in thousands of textbooks and scientific publications; this in contrast to other natural science domains such as physics or chemistry, where knowledge is to a large extent formalized, often in mathematical terms. This is in part connected to the enormous complexity of the studied objects: a biological cell, tissue and organism, under normal or pathological conditions. In order to facilitate reviewing and navigating through the huge and complex corpus of available knowledge of molecular cell biology, it is tempting to consider representing it in the form of a large hierarchically organized map. For long such an effort has not been achievable since our knowledge of biology has been too fragmented, and we have been lacking a seamless representation of

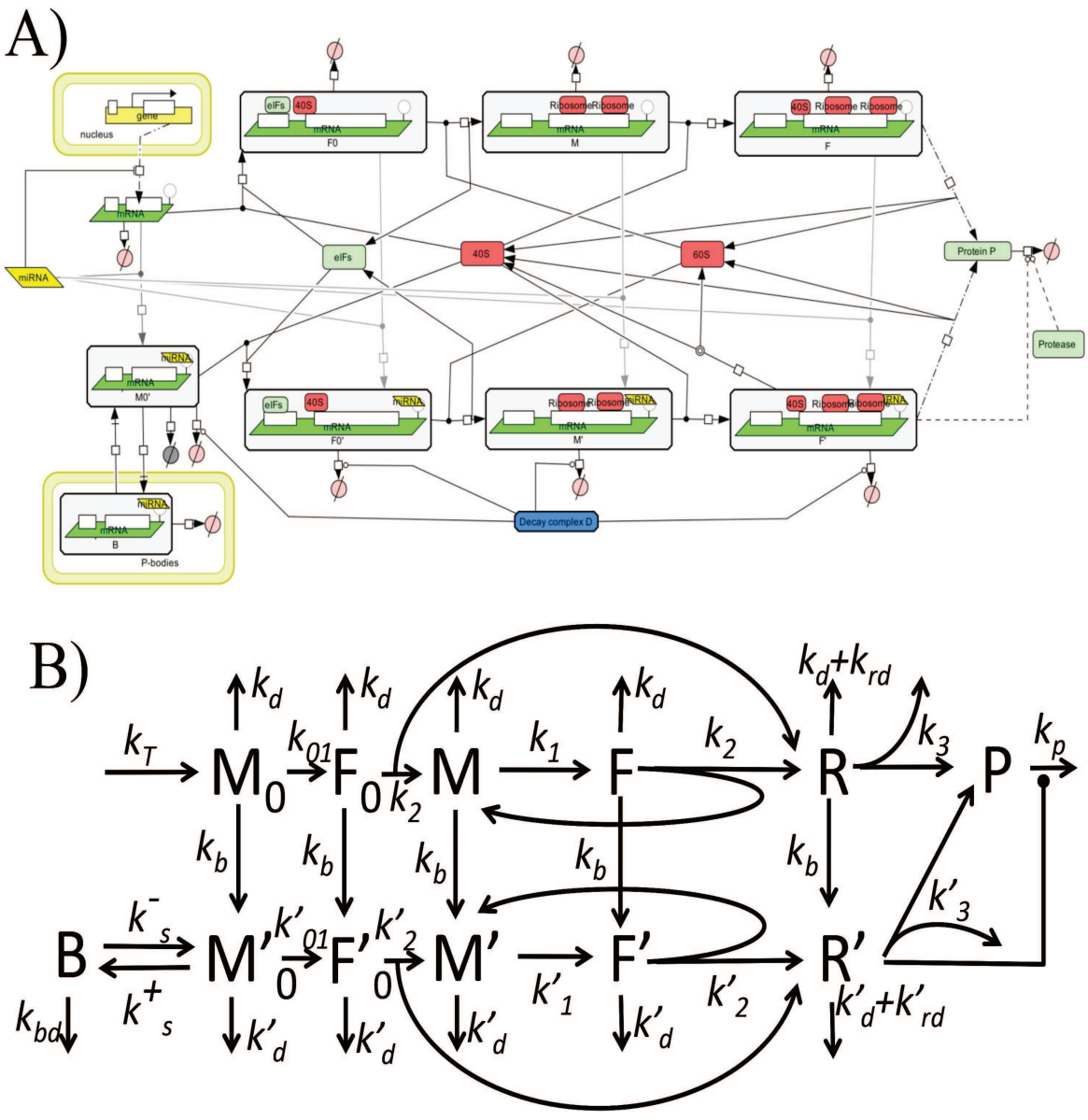

FIGURE 4. Mathematical model of miRNA-mediated mechanisms of translation repression. This model combines 9 known mechanisms in one reaction network. A) SBGN-based representation of the model. B) Schematic view of the reaction network where M0 is newly produced mRNA, F0 - mRNA initiated with 40S ribosome component for the first round of translation, M and F - are amounts of translated mRNAs without or with 40S ribosomal component sitting at the translation initiation site, R - is the total amount of translating ribosomes and P is the amount of protein. Prime symbol denotes the corresponding states of mRNA with miRNA bound. The derivation of the model equations is described in [26].

the molecular processes taking place in a cell. As a consequence, first efforts to organize this knowledge presented it as a juxtaposition of unconnected pieces of information, like plethora of pathway maps, a presentation which might falsely suggest that they are independent or loosely interconnected. Although the consequences of this bias is still visible today in the research practice and in the resources describing our knowledge, we think that the scientific community has now reached enough understanding to undertake the reconstruction of a global map of molecular cell biology. This map will be neither complete nor error-free, but by its holistic nature should open new perspectives.

Charting maps, starting from geography to scientific graphics, which also stemmed from cartesian geography, is an essential scientific activity in various fields. Maps do not only chart the territory but also our understanding [21]. In molecular biology, diagrams drawing fragments of various molecular biology mechanisms is a commonly used way to map and recapitulate our knowledge.

In 1960s, an impressive effort has been undertaken to collect many pieces of the knowledge on cellular metabolism into a global metabolic map, which became one of the most used resource in the field [17]. Importantly, the metabolic map reflects not only local relations between metabolites but also represents our understanding of higher-level relations between metabolic functions, using carefully designed meaningful global layout and principles of semantic zooming [60]. In this sense, the information on the metabolic map is not equivalent to a list of biochemical reactions represented in it, but also visualizes the proximity of functions. In late 90s, several large maps related to mammalian cell cycle and cell fate decisions have been charted using an original diagrammatic technique, also providing an insightful map layout [47].

Efforts in systems biology to have a common standard graphical language (SBGN) to depict the biological knowledge [57], and handful software for charting the large biological maps (CellDesigner [46], CellIllustrator and others) and systems biology tools for analysing and manipulating them [5,6,49,65,76] stimulated the community to create large comprehensive maps of molecular interactions implicated in various biological processes. In this way, large maps of signaling pathways (such as the map of RB/E2F pathway [8]), complex metabolic processes, maps specific for certain cell types or non-mammalian organisms appeared. Several comprehensive maps representing biological mechanisms that are implicated and rewired in particular diseases such as Alzheimer disease or rheumatoid arthritis have been also recently charted (a comprehensive collection of these maps can be found at http://navicell.curie.fr).

In cancer biology, the idea of representing molecular processes involved in tumorigenesis in the form of a complex map was proposed more than a decade ago. Weinberg and Hanahan systematized available knowledge on the signalling cascades implicated in cancer and drafted the famous simplified diagram of cancer signalling with an image of a human cell in the background [42, 43]. To promote this idea, the authors later suggested using a metaphor of subway map with "stations" corresponding to cancer driver genes and the lines corresponding to cancer pathways [41]. Several attempts have been undertaken to connect cancer driver genes into relatively large and comprehensive networks or represent them as pathway databases such as Cancer Cell Map as a part of Pathway Commons database [13].

It is tempting to imagine a global map of molecular biology, at least for the most important organisms or diseases, by merging and integrating some of the existing maps together. However, without proper technology for navigating, exploration and construction of very large maps, this effort is meaningless and the maps would not be useful given their size and complexity. Nowadays, computer technologies give an opportunity to browse very large and complex geographical maps intuitively, using user-friendly web-interfaces such as Google Maps. The influence of such technologies on molecular biology and systems biology has been modest until today. Nevertheless, many pathway browsers reproduce elements of the Google Maps interface and some of them already use this technology directly to browse CellDesigner maps [20, 49]. Effective exploration of maps requires the development of methods for abstracting the network information in order to visualize different level of detail, depending on the level of zooming (such as semantic zooming [52,60]). Providing adaptive network zoom views is a critical under-developed component in most of the existing tools.

Recently we developed a scalable methodology allowing the assembly of very large comprehensive maps of molecular interactions and browsing them using the Google Maps technology and the princi-

ples of semantic zooming [7, 49]. Using this methodology, we charted the largest (to our knowledge) comprehensive map focusing on major molecular mechanisms implicated in cancer, the Atlas of Cancer Signalling Networks (ACSN), freely available online at http://acsn.curie.fr (Figure 5). We present this map in the form of a web-based atlas which is hierarchical and interconnected collection of maps browsable online [50]. The atlas depicts molecular mechanisms of Cell Cycle [8], DNA Repair, Cell Survival, Apoptosis, Epithelial-to-Mesenchymal Transition and Cell Motility and beyond. ACSN already describes functioning of the majority (60%) of cancer driver genes as they were defined recently in [66], and we constantly enlarge the maps with new data. The global ACSN map is decomposed to tens of submaps more conveniently representing smaller-scale molecular processes. ACSN is a working tool applicable in cancer systems biology, which will be extended in the near future with other relevant biological processes in order to cover all cancer hallmarks. ACSN can be used as an ordinary pathway database, but due to its concept of being a global map of cancer molecular biology, it provides also novel opportunities for exploring molecular and interaction information. For example, it allows mapping quantitative data (such as transcriptome, copy number profiles, mutational profiles) on top of it (Figure 5) [7]. The content of ACSN is original and not a mere extraction from other pathway databases, which makes it a valuable pathway resource per se. We demonstrated by quantitative comparison that, in the area of its scope, ACSN collects more up-to-date interaction information than comparable existing pathway databases. In addition, ACSN provides a discussion forum in the form of a weblog for commenting the content of ACSN by its users.

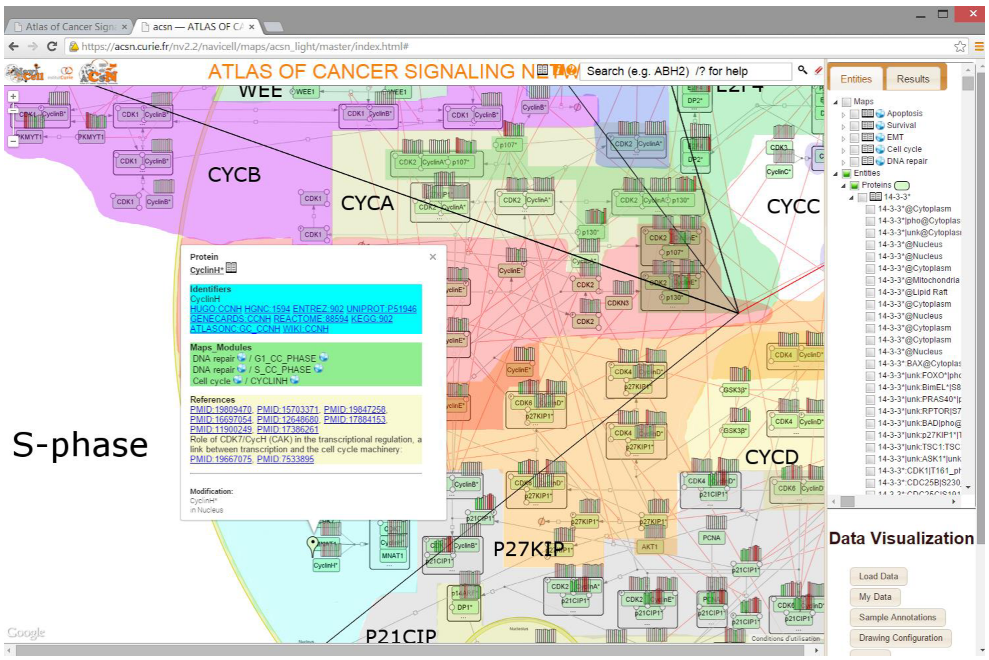

FIGURE 5. A screenshot of the cell cycle territory in the map of Atlas of Cancer Signalling Network (ACSN, http://acsn.curie.fr), with profiles of expression in several tumour samples shown on top of the protein icons [7, 50].

Formalization of knowledge on the functioning of biological networks is an important step in dealing with their complexity. Using comprehensive (not over-simplified) reconstructions of cancer-related processes, we can have a realistic idea on the complexity of their structure, and start developing ade-

quate algorithms for quantifying and abstracting this complexity for further use in data analysis and mathematical modeling of the kind presented below.

4.2. Cell fate decisions in cancer cells

The popular Einstein's saying that God integrates empirically can be repeated with respect to the biological cells too. A biological cell constantly performs complex integration of the signals coming from the environment. This integration is also conditioned on the internal cell state. The result of such "integration" can be "a value" from a discrete set, i.e. a decision to initiate and launch one of the existing cellular programs. For example, a cell can trigger the apoptotic or necroptotic program, i.e. a program of cellular suicide, as a response to some unfavorable conditions (e.g., irreparable DNA damage or starvation) or as a part of the normal development program. Another example of such a decision - triggering Epithelial-To-Mesenchymal Transition (EMT) - is a requirement for cellular motility. Let us call "cell fates" to name triggering such programs (some of them are irreversible), and "cell fate decision process" to name the process of "empirical integration".

The cell fate decisions are computed by the cells through complex signal transduction pathways and their interactions. The design of these pathways is a result of optimization by natural selection of two mutually opposite requirements: a possibility to control the decisions, and robust functioning, i.e. resistance to various perturbations, including mutations in the pathway genes [1]. Certain mutations or combinations of them can drastically change the cell fate decision process which can lead to various diseases.

One of the most crucial cell fate decision is a decision between cell survival and death. Tumour initiation and progression is characterised by a violation of a balance between cell survival and cell death, which is tightly controlled in normal conditions, towards excessive survival in cancer. The balance is normally ensured by a system of molecular switches that trigger irreversible cellular decisions for a particular cell fate in some particular conditions. Compromising the normal function of these switches leads to carcinogenesis and other systemic diseases.

For systems biology of cancer, a very important task is to reveal the logics of the natural empirical integrations, and reproduce it, to some extent, on computer in the form of mathematical models. We found out that it is convenient to describe the process of cell fate decisions by discrete (logical) modeling which has been developed in computational biology since 1960s [3]. The process of cellular fate decisions in this approach is recapitulated in the form of a state transition graph, describing a finite number of cellular states and transitions between them characterized by certain probabilities. Some of the cellular states represent fixed points, and they are associated with cell fates (or phenotypes). Studying random walks along the state transition graphs, it is possible to estimate the probability of reaching particular fixed points which can be interpreted as a probability to observe a particular cell fate in the biological experiment. The state transition graphs are usually represented in a compact fashion using a biological regulatory network (a directed graph). The nodes (molecules, their modifications, or molecular processes) in this network are assumed to be characterized by a discrete number (state), accompanied by the rules of updating this number based on the states of their immediate upstream neighbors (regulators). In the simplest case, the state is "0" or "1", representing an "active" or "inactive" state.

We used this approach to recapitulate the process of cell survival-cell death decisions in a well-known experiment where the cells are exposed to a concentration of TNF of FASL proteins [9]. As a result, the cells can induce necrotic or apoptotic programmed cell death or survive through induction of NFkB pathway (see Figure 6). The most reliable knowledge about functioning of the cell fate decision molecular machinery was assembled, using GINsim software [16], in the form of a regulatory network (Figure 6a), where the nodes represents proteins or their modifications, small molecules (such as ROS) and molecular processes (such as Mitochondrial Permeability Transition (MPT)). Each node was accompanied by an update rule (for example, how BAX protein changes its state depending on the states of CASP8 and BCL2). The state transition graph was computed and analyzed for probability of observing Survival,

Necrosis, Apoptosis cell fates in the wild-type cell, and in all single mutants where the state of certain proteins was fixed to "inactive" or "active" states (Figure 6).

These predictions were systematically compared with the experimental data on the cell death phenotypes observed in various mutant experimental systems, including cell cultures and mice (Figure 6c, [9]). The model was able to qualitatively recapitulate all of them and to suggest some new yet unexplored experimentally mutant phenotypes. The most interesting in this setting would be to consider synthetic interactions between individual mutants, when several nodes on the diagram are affected by a mutation.

We developed a toolbox of methods to analyze some particular properties of the discrete model. For example, we were able to predict cell fate decisions after transient exposure to ligands. We tested importance of the "switching speed" of certain proteins and importance of individual regulatory links for the process of cell fate decision [75]. We used specific model reduction techniques to reduce the model to a minimal configuration which revealed the pattern of its organization: "a tri-stable switch" (Figure 6d, [9, 11, 70]). Recently, we developed a computational method for predicting genetic interactions from Boolean models [10].

In another project, we used this methodology to predict the effect of mutations in the regulatory mechanism which controls the switch between epithelial and mesenchymal cell phenotypes. This switch (EMT) is associated with appearance of metastases in cancer, because it allows the cells to detach from the primary tumor and become mobile in order to arrive to the distant organs. The regulatory mechanism was assembled from the P53, WNT and NOTCH pathways whose involvement in EMT regulation was demonstrated (the description is available at https://navicell.curie.fr/navicell/maps/notchp53wnt/master/). We predicted that rather than a single mutation alone, a combination of P53 knock-out and overexpression of NOTCH would induce EMT. This prediction was validated in a genetically modified mouse model of intestinal cancer, where, indeed, conditional overexpression of NOTCH and downregulation of P53 lead to the induction of EMT at the front of the primary tumor and to rapid and early metastasation, which was not possible to achieve before [15, 51].

Another example of cell fate decision modeling is predicting the ability of a cell to repair its damaged DNA. One of the most dangerous defects of DNA is its double-strand break. Left unrepaired, it is usually lethal for the cell. One of the DNA repair pathways capable to repair the double-strand breaks is already above-mentioned Homologous Recombination (HR) pathway. Some double knockouts of the genes involved in HR are lethal (for example, knocking out Srs2 and Rad54 [58]). This observation is intriguing and counter-intuitive for two reasons. First of all, the HR pathway is not an essential pathway in yeast and can be compensated by other pathways. Second, synthetic-lethality between genes in the same pathway does not follow the classical paradigm of the between-redundant-pathway model. Studying this question, it was noticed that some of the transitions between repairing DNA states are reversible and regulated: for example, Srs2 regulates one of the backward transitions. This led to the idea of explaining this synthetic lethality by the "kinetic trap" model. According the model, due to the effect of two mutations a pathway is trapped in an intermediate state which might be toxic for the cell. This situation was analyzed formally, exploiting mathematical modeling, and the quantitative aspects of this mechanism were clarified [72]. Moreover, it was hypothesized that the kinetic trap mechanism can be responsible for appearance of synthetic lethality in other molecular processes, not necessarily related to DNA repair. For example, kinetic trapping can make an essential cascade of reversible post-translational protein modifications being trapped in a state not allowing the signal to propagate.

All these examples shows that we can deal with relatively complex models of molecular mechanisms and successfully create tools to predict and manipulate biological systems towards the desired behavior despite their non-intuitive complex dynamics.

5. Conclusion and perspectives

Working with problems of biological complexity has led me to a rather optimistic conclusion: the complexity of biological systems, as they are represented by formal mathematical models, can be tackled [68].

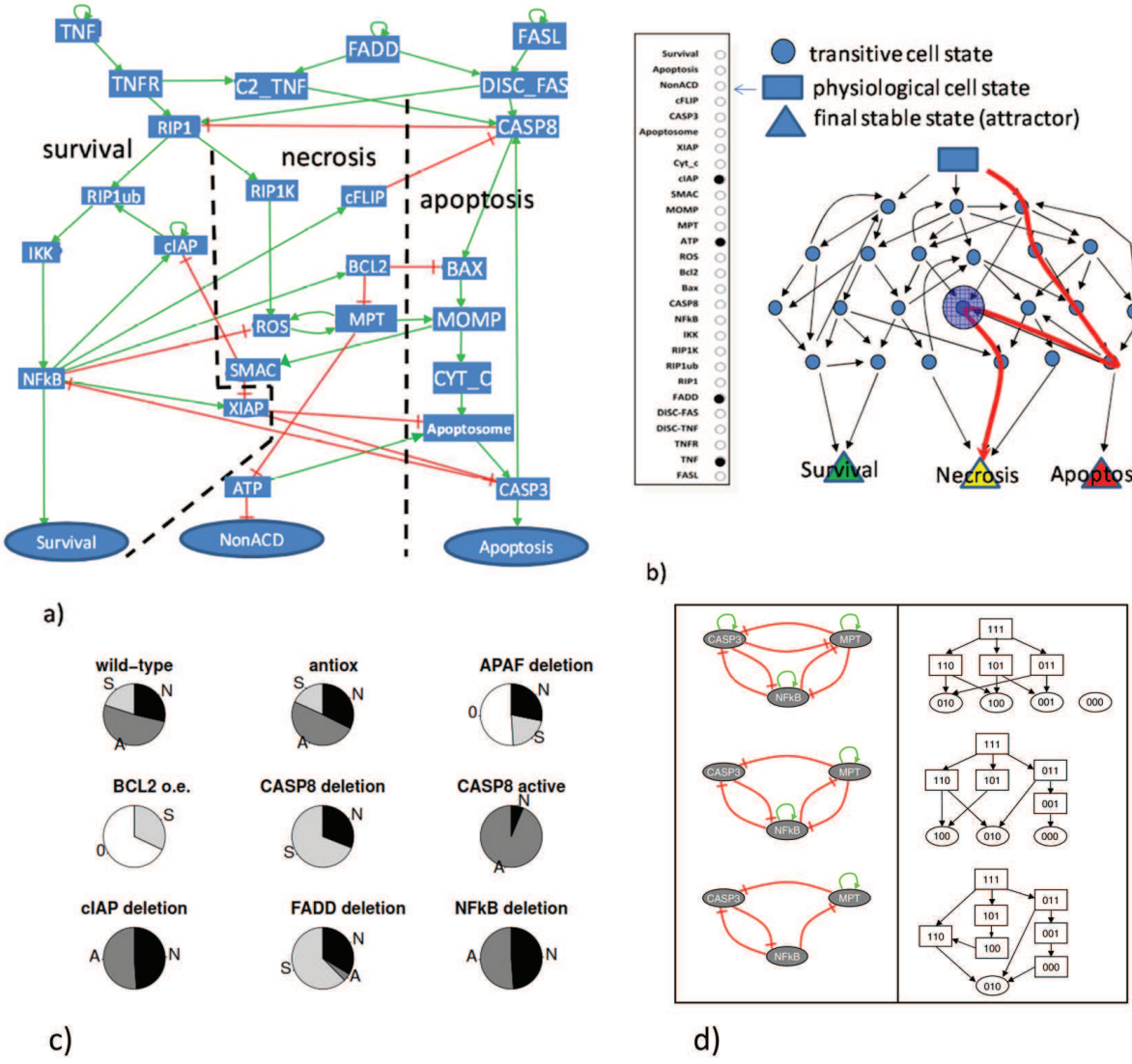

FIGURE 6. Logical model of cell fate decisions between survival, apoptosis and programmed necrosis (necroptosis). Here "A" denotes Apoptosis, "N" denotes Necrosis and "S" denotes Survival, "0" denotes Naive state, in which the cell is not inivolved in the cell fate decision process. a) The diagram describing the cell fate decision mechanisms; b) cartoon illustration of a state transition graph on how the probabilities of different cell fates are computed; c) distribution of probabilities of cell fates for some mutant simulations; d) "tri-stable trigger" pattern of the cell fate decision mechanism (top row) with the corresponding state transition graph (on the right) together with various mutation types. Reproduced from [9, 70].

Their complexity is frequently only seaming, representing rather complicatedness of the system, its ability to function in many different situations. However, each particular system's response to a perturbation can be simple and comprehensible. Thus, the complexity can be reduced by dissection or simplification, provided appropriate experimental and computational methods.

This conclusion is a pleasant surprise rather than an expectation: one could have assumed that the natural selection has tendency to "complexify" biological systems, pushing them to the "wild" complexity situation. My conclusion can be misleading as well, and related only to our way of representing the biological reality by experimental data or mathematical models. It might happen that these models or data are not adequate to grasp the essential complexity of a living cell. However, I do not think so.

Of course, we are very far from feeling comfortable being faced with the biological complexity. We lack approaches allowing working and simulating large and incompletely characterized biochemical networks. We do not know how to systematically integrate the existing data in the most useful way. Most importantly, we do not know how to use the available data in the most efficient way using our knowledge about cellular molecular mechanisms, how to efficiently produce new knowledge on the functioning of molecular mechanisms from the large-scale genomic-wide data. Lack of good and objectively validated approaches in these fields is reflected in difficulties with realistic and large-scale modeling of cells, tissues, organs, organisms.

Nevertheless, the progress in quantifying, understanding and reducing complexity of biological systems is real and inevitable. Recent breakthroughs in multi-level profiling of molecular composition of a cell, whole-cell computational models, and realistic multi-scale models describing both biochemical and biophysical properties of cell populations make this direction of scientific research exciting and full of promises.

Acknowledgements. I am thankfull to Alexander Gorban, my scientific supervisor and the "scientific father" for his generosity in sharing ideas, his friendship and support. I thank all past and present members of "Computational Systems Biology of Cancer" group (http://sysbio.curie.fr) whose projects I coordinate for ten years together with Emmanuel Barillot whom I thank for all good advises and supporting many of my initiatives. Various projects mentioned in this study have been supported by funding from internal projects of Institut Curie, European Union FP7 program (APO-SYS and ASSET projects), by French National Cancer Institutes (ITMO Cancer and INCA) and French National Research Agency (ANR).

References

[1] E. Barillot, L. Calzone, P. Hupe, J.P. Vert, A. Zinovyev. Computational systems biology of cancer. Chapman & Hall, CRC Mathematical & Computational Biology, 2012.

[2] S.C. Bendall, K.L. Davis, A.D. el Amir, M.D. Tadmor, E.F. Simonds, T.J. Chen, D.K. Shenfeld, G.P. Nolan, D. Pe'er. *Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.* Cell, 157 (3) (2014), 714–325.

[3] D. Bérenguier, C. Chaouiya, P.T. Monteiro, A. Naldi, E. Remy, D. Thieffry, L. Tichit. *Dynamical modeling and analysis of large cellular regulatory networks.* Chaos, 23 (2) (2013), 025114.

[4] A. Biton, I. Bernard-Pierrot, Y. Lou, C. Krucker, E. Chapeaublanc, P. Rubio, B. Lopez, A. Kamoun, Y. Neuzillet, P. Gestraud, G. Grieco, S. Rebouissou, A. de Reynies, S. Benhamou, T. Lebret, J. Southgate, E. Barillot, Y. Allory, A. Zinovyev, F. Radvanyi. *Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes.* Cell Reports, 9 (2014), 1–11.

[5] E. Bonnet, L. Calzone, D. Rovera, G. Stoll, E. Barillot, A. Zinovyev. *BiNoM 2.*0, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. BMC Syst. Biol., 7 (1) (2013), 18.

[6] E. Bonnet, L. Calzone, D. Rovera, G. Stoll, E. Barillot, A. Zinovyev. *Practical use of BiNoM: a Biological Network Manager Software.* Methods Mol. Biol., 1021 (2013), 127–146.

[7] E. Bonnet, E. Viara, I. Kuperstein, L. Calzone, D.P.A. Cohen, E. Barillot. *NaviCell Web Service for network-based data visualization.* Nucleic Acids Research (2015), Advanced Access Publication, http://dx.doi.org/10.1093/nar/gkv450.

[8] L. Calzone, A. Gelay, A. Zinovyev, F. Radvanyi, E. Barillot. *A comprehensive modular map of molecular interactions in RB/E2F pathway.* Mol. Syst. Biol., 4 (2008), 174.

[9] L. Calzone, L. Tournier, S. Fourquet, D. Thieffry, B. Zhivotovsky, E. Barillot, A. Zinovyev. *Mathematical modelling of cell-fate decision in response to death receptor engagement.* PLoS Comput. Biol., 6 (3) (2010), e1000702.

[10] L. Calzone, E. Barillot, A. Zinovyev. *Predicting genetic interactions from Boolean models of biological networks*. Integrative Biology (2015), Advanced Access Publication, http://pubs.rsc.org/en/Content/ArticleLanding/2015/IB/C5IB00029G.

[11] L. Calzone, A. Zinovyev, B. Zhivotovsky. Understanding Different Types of Cell Death Using Systems Biology. In Systems Biology of Apoptosis (ed. by Lavrik, I.). Springer, 2012.

[12] A. Carbone, A. Zinovyev, F. Kepes. *Codon Adaptation Index as a measure of dominating codon bias*. Bioinformatics, 19 (13) (2003), 2005–2015.

[13] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G.D. Bader, C. Sander. *Pathway Commons, a web resource for biological pathway data*. Nucleic Acids Res., 39 (Database issue) (2011), D685–D690.

[14] M. Chacòn, M. Lévano, H. Allende, H. Nowak. *Detection of gene expressions in microarrays by applying iteratively elastic neural net*. In: B. Beliczynski et al. (Eds.). Lecture Notes in Computer Sciences, Springer: Berlin–Heidelberg, 4432 (2007), 355–363.

[15] M. Chanrion, I. Kuperstein, C. Barrière, F. El Marjou, D. Cohen, D. Vignjevic, L. Stimmer, P. Paul-Gilloteaux, I. Bièche, R. Tavares Sdos, G.F. Boccia, W. Cacheux, D. Meseure, S. Fre, L. Martignetti, P. Legoix-Né, E. Girard, L. Fetler, E. Barillot, D. Louvard, A. Zinovyev, S. Robine. *Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut*. Nature Communications, 5 (2014), 5005.

[16] C. Chaouiya, A. Naldi, D. Thieffry. *Logical modelling of gene regulatory networks with GINsim*. Methods Mol. Biol., 804 (2012), 463–479.

[17] S. Dagley, D. Nicholson. An introduction to metabolic pathways. Blackwell Scientific Publications, 1970.

[18] Y. Drier, M. Sheffer, E. Domany. *Pathway-based personalized analysis of cancer*. PNAS, 110 (16) (2013), 6388–6393.

[19] H. Failmezger, B. Jaegle, A. Schrader, M. Hülskamp, A. Tresch. *Semi-automated 3D leaf reconstruction and analysis of trichome patterning from light microscopic images*. PLoS Comp. Biol., 9 (4) (2013), e1003029.

[20] L.A. Flórez, C.R. Lammers, R. Michna, J. Stülke. *CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling and regulatory pathways*. Bioinformatics, 26 (23) (2010), 2997–2999.

[21] S.H. Friend, T.C. Norman. *Metcalfe's law and the biology information commons*. Nat. Biotechnol., 4 (2013), 297–303.

[22] A.N. Gorban. Multigrid Integrators on Multiscale Reaction Networks. Keynote talk given at Algorithms for Approximation VI, Ambleside, the Lake District, UK, 2009.

[23] A.N. Gorban, I. Karlin, A. Zinovyev. *Constructive Methods of Invariant Manifolds for Kinetic Problems*. Physics Reports, 396 (2004), 197–403.

[24] A.N. Gorban, I. Karlin, A. Zinovyev. *Invariant grids for reaction kinetics*. Physica A, 333 (2004), 106–154.

[25] A.N. Gorban, B. Kegl, D. Wunch, A. Zinovyev. (eds.) Principal Manifolds for Data Visualisation and Dimension Reduction. Lecture Notes in Computational Science and Engeneering 58, 2008.

[26] A.N. Gorban, N. Morozova, A. Harel-Belan, A. Zinovyev. *Basic and simple mathematical model of coupled transcription, translation and degradation*. (2013) http://arxiv.org/abs/1204.5941.

[27] A.N. Gorban, O. Radulescu, A.Y. Zinovyev. *Asymptotology of chemical reaction networks*. Chem. Eng. Sci., 65 (2010), 2310–2324.

[28] A.N. Gorban, N. Sumner, A. Zinovyev. *Topological grammars for data approximation*. Appl. Math. Lett., 20 (4) (2007), 382–386.

[29] A.N. Gorban, N. Sumner, A. Zinovyev. *Beyond The Concept of Manifolds: Principal Trees, Metro Maps, and Elastic Cubic Complexes*. Lecture Notes in Computational Science and Engeneering 58 (2008), 223–240.

[30] A.N. Gorban, G.S. Yablonsky. *Grasping Complexity*. Computers & Mathematics with Applications, 65 (10) (2013), 1421–1426.

[31] A.N. Gorban, A.Y. Zinovyev. *Method of Elastic Maps and its Applications in Data Visualization and Data Modeling*. International Journal of Computing Anticipatory Systems, Chaos 12 (2001), 353–369.

[32] A.N. Gorban, A. Zinovyev. *Elastic Principal Graphs and Manifolds and their Practical Applications*. Computing, 75 (2005), 359–379

[33] A.N. Gorban, A. Zinovyev. *Elastic Maps and Nets for Approximating Principal Manifolds and Their Application to Microarray Data Visualization*. Lecture Notes in Computational Science and Engeneering, 58 (2008), 97–128.

[34] A.N. Gorban, A.Y. Zinovyev. Principal Graphs and Manifolds. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques (eds. Olivas E.S., Guererro J.D.M., Sober M.M., Benedito J.R.M., Lopes A.J.S.). Information Science Reference, 2009.

[35] A.N. Gorban, A. Zinovyev. *Principal manifolds and graphs in practice: from molecular biology to dynamical systems*. Int. J. Neural Syst., 20 (3) (2010), 219–232.

[36] A.N. Gorban, A.Y. Zinovyev, A.A. Pitenko. *Visualization of data using method of elastic maps* (in Russian). Informatsionnie technologii, (6) (2000), 26–35.

[37] A.N. Gorban, A.Y. Zinovyev, T.G. Popova. *Seven clusters in genomic triplet distributions*. In Silico Biology, 3 (4) (2003), 471–482. http://arxiv.org/abs/cond-mat/0305681 [cond-mat.dis-nn]

[38] A.N. Gorban, A. Zinovyev, D.C. Wunsch. *Application of the method of elastic maps in analysis of genetic texts*. In: Proceedings of International Joint Conference on Neural Networks (IJCNN2003). Portland, Oregon, Vol. 3, 2003, 1826–1831.

[39] M. Gromov. Metric Structures for Riemannian and Non-Riemannian Spaces. Progress in Mathematics 152. Birkhauser Verlag, 1999.

[40] M. Gromov. Allure of Quotations and Enchantment of Ideas, 2013. http://www.ihes.fr/∼gromov/PDF/quotationsideas.pdf.

[41] W.C. Hahn, R.A. Weinberg. *A subway of cancer pathways*. Nature Reviews Cancer Poster, (2002).

[42] D. Hanahan, R.A. Weinberg. *The hallmarks of cancer*. Cell, 100 (1) (2000), 57–70.

[43] D. Hanahan, R.A. Weinberg. *Hallmarks of cancer: the next generation*. Cell, 144 (5) (2011), 646–674.

[44] W.S. Hlavacek. *How to deal with large models?* Mol. Syst. Biol., 5 (2009), 240.

[45] E.M. Izhikevich, G.M. Edelman. *Large-scale model of mammalian thalamocortical systems*. PNAS, 105 (2008), 3593–3598.

[46] H. Kitano, A. Funahashi, Y. Matsuoka, K. Oda. *Using process diagrams for the graphical representation of biological networks*. Nat. Biotechnol., 23 (8) (2005), 961–966.

[47] K.W. Kohn. *Molecular interaction map of the mammalian cell cycle control and DNA repair systems*. Mol. Biol. Cell., 10 (8) (1999), 2703–2734.

[48] M.D. Kruskal. *Asymptotology*. In: Dobrot, S. (Ed.), Mathematical Models in Physical Sciences. Prentice-Hall, Englewood Cliffs, NJ, (1963), 17–48.

[49] I. Kuperstein, D.P. Cohen, S. Pook, E. Viara, L. Calzone, E. Barillot, A. Zinovyev. *NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps*. BMC Syst. Biol., 7 (2013), 100.

[50] I. Kuperstein, L. Grieco, D.P.A. Cohen, D. Thieffry, A. Zinovyev, E. Barillot. *The shortest path is not the one you know: application of biological network resources in precision oncology research*. Mutagenesis 30 (2015), 191–204.

[51] I. Kuperstein, S. Robine, A. Zinovyev. *Computational biology helps finding genetic determinants of metastatic colon cancer*. Cell Cycle (2015), In press.

[52] M. Latendresse, P.D. Karp. *Web-based metabolic network visualization with a zooming user interface*. BMC Bioinformatics, 12 (2011), 176.

[53] G. Mathonnet et al.. *MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F*. Science, 317 (2007), 1764–1767.

[54] J.H. Miller, S.E. Page. Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton University Press, 2007.

[55] N. Morozova, A. Zinovyev, N. Nonne, L.L. Pritchard, A.N. Gorban, A. Harel-Bellan. *Kinetic signatures of microRNA modes of action*. RNA, 18 (9) (2012), 032284.

[56] T. Nissan, R. Parker. *Computational analysis of miRNA-mediated repression of translation: implications for models of translation initiation inhibition*. RNA, 4 (8) (2008), 1480–1491.

[57] N. Le Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M.I. Aladjem, S.M. Wimalaratne, F.T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S.E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T.C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D.B. Kell, C. Sander, H. Sauro, J.L. Snoep, K. Kohn, H. Kitano. *The Systems Biology Graphical Notation*. Nat. Biotechnol., 27 (8) (2009), 735–741.

[58] F. Palladino, H.L. Klein. *Analysis of mitotic and meiotic defects in Saccharomyces cerevisiae SRS2 DNA helicase mutants*. Genetics, 132 (1992), 23–37.

[59] K. Pearson. *On lines and planes of closest to systems of points in space*. Philos. Mag., 2 (1901), 559–572.

[60] S. Pook, G. Vaysseix, E. Barillot. *Zomit: biological data visualization and browsing*. Bioinformatics, 14 (9) (1998), 807–814.

[61] O. Radulescu, A.N. Gorban, S. Vakulenko, A. Zinovyev. *Hierarchies and modules in complex biological systems*. In: Proceedings of European Conference on Complex Systems. Oxford, UK, 2006.

[62] O. Radulescu, A.N. Gorban, A. Zinovyev, A. Lilienbaum. *Robust simplifications of multiscale biochemical networks*. BMC Syst. Biol., 2 (2008), 86.

[63] O. Radulescu, A.N. Gorban, A. Zinovyev. *Reduction of dynamical biochemical reactions networks in computational biology*. Frontiers in Genetics, 3 (2012), 00131.

[64] O. Radulescu, A. Zinovyev, A. Lilienbaum. *Model reduction and model comparison for NFkB signalling*. In Proceedings of Foundations of Systems Biology in Engineering, Stuttgart, Germany, (2007).

[65] P. Vera-Licona, E. Bonnet, E. Barillot, A. Zinovyev. *OCSANA: Optimal Combinations of Interventions from Network Analysis*. Bioinformatics, 15 (29) (2013), 1571–1573.

[66] B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz, K.W. Kinzler. *Cancer genome landscapes*. Science, 339 (2013), 1546–1558.

[67] A. Wagner. Robustness and Evolvability in Living Systems. Princeton Univ. Press, 2005.

[68] A. Zinovyev. *Dealing with complexity of biological systems: from data to models*. HDR synthesis text, (2014). http://arxiv.org/abs/1404.1626.

[69] A. Zinovyev. Visualization of Multidimensional Data (in Russian). KGTU Publ., Krasnoyasrk, 2000.

[70] A. Zinovyev, S. Fourquet, L. Tournier, L. Calzone, E. Barillot. *Cell death and life in cancer: mathematical modeling of cell fate decisions*. In Advances in Experimental Medicine and Biology (Goryanin, I. and Goryachev A, eds.), Springer, 736 (2012), 682.

[71] A. Zinovyev, U. Kairov, T. Karpenyuk, E. Ramanculov. *Blind Source Separation Methods For Deconvolution Of Complex Signals In Cancer Biology*. Biochemical and Biophysical Research Communications, 430 (3) (2013), 1182–1187.

[72] A. Zinovyev, I. Kuperstein, E. Barillot, W.-.D. Heyer. *Synthetic Lethality between Gene Defects Affecting a Single Non-essential Molecular Pathway with Reversible Steps*. PLoS Comput. Biol., 9 (4) (2013), e1003016.

[73] A. Zinovyev, E. Mirkes. *Data complexity measured by principal graphs*. Computers and Mathematics with Applications, 65 (2013), 1471–1482.

[74] A. Zinovyev, N. Morozova, A.N. Gorban, A. Harel-Belan. *Mathematical modeling of microRNA-mediated mechanisms of translation repression*. In MiRNA Cancer Regulation: Advanced Concepts, Bioinformatics and Systems Biology Tools (Schmitz U, Wolkenhauer O, Vera J, eds.), Springer, (2013), 189–224.

[75] A. Zinovyev, N. Morozova, N. Nonne, E. Barillot, A. Harel-Bellan, A.N. Gorban. *Dynamical modeling of microRNA action on the protein translation process*. BMC Syst. Biol., 4 (2010), 13.

[76] A. Zinovyev, E. Viara, L. Calzone, E. Barillot. *BiNoM: a Cytoscape plugin for using and analyzing biological networks*. Bioinformatics, 24 (6) (2008), 876–877.