Phenotype State Spaces and Strategies for Exploring Them

Andreas Hadjiprocopis[1] and Rune Linding[2]

[1] The Institute of Cancer Research (ICR), SW3 6JB, London, UK
[2] Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), DK-2800 Lyngby, Denmark

0.1 Introduction

Proteins and their interactions determine how cells behave. Genes are the blueprints for protein synthesis; their activation or suppression determines the absence or presence of a protein which in turn can give rise to further activation or suppression of other genes and proteins. This chemical chain reaction usually involves positive and negative feedback loops and is subjected to stochastic noise and the influence of environmental factors. Moreover, epistasis - the cancellation or modification of a gene's contribution to the phenotype by other genes - is generally the rule rather than the exception in genetics.

Despite all these factors, amazingly robust cell behavior is apparent in many biological systems although it is difficult to be modelled and/or predicted. Building the topology and quantifying the direct and indirect cause-effect (stimulus, expression, activation, behavior) relationships of the reactions leading to the phenotypes - in general, genetic regulatory networks (GRN) - is challenging in at least three ways.

Firstly, how are these relationships described? Traditionally, mathematical models are expressed in terms of transfer functions relating inputs to outputs expressed as a composition of differential equations with a time dimension. We argue though that the cellular signaling networks are probabilistic in nature and that diffusion based models remain challenging due to lack of knowledge of essential system parameters, such as rate constants. Most importantly, treating intracellular protein and gene interactions as in-vitro chemical reactions might not be safe because the usual assumption of diffusion dynamics namely that of the free movement of a sufficiently large number of molecules is usually the exception rather than the rule due to the very small number of reactant molecules in highly confined and crowded space. Moreover, the concentration of a protein is highly dependent on sub-cellular localization and thus the picture of the cell as a homogeneous mix container is simply wrong. Of even more profound impact, many signaling systems are centered on or around scaffolding proteins mimicking solid state chemical environments and has little or no resemblance to diffusion limited systems. For these reasons, significant statistical fluctuations in the behavior of the reactions are common and the use of diffusion limited modelling approaches not appropriate.

More recently, computational models in the form of Boolean networks, Petri nets, Interacting (Probabilistic) State Machines have been proposed (Fisher & Henzinger 2007) as an alternative way of modeling biological processes with the added benefits of abstraction, high-level reasoning and mature process calculus tools to aid analysis.

Secondly, our ability to gain understanding and abstract from derived molec-

Acknowledgment: AH wishes to thank Dr Herbert Wiklicky for the useful comments he made on the manuscript and during their many discussions.

ular mechanisms and regulatory networks topology is really limited because of the sheer complexity and abundance of low level information inherent in these networks. In this respect, a way to reduce complexity is by treating the parts of these networks where the interactions within are much larger compared to the interactions between, as semi-autonomous modules - essentially, as black boxes. Modular response analysis (Bruggeman, Westerhoff, Hoek & Kholodenko 2002) offers such a framework. Building on this, (Zamir & Bastiaens 2008) developed a methodology for reverse engineering network structure in order to analyse how perturbations propagate in a network. Modularisation at a higher level is also key for reusing parts of the derived regulatory networks.

Finally, our ability to understand GRN networks is impacted by the fact that, kinase activation can initiate different cellular decisions depending on the pre-activation state of the network. In (Janes, Albeck, Gaudet, Sorger, Lauffenburger & Yaffe 2005), it was shown that JNK activation can be anti- or pro-apoptotic depending on network state when cells received growth factor cues. Therefore, to describe and predict a cellular response to a perturbation, studies must be carried out in the context of the cell's multivariate network state (Linding 2010).

0.2 Phenotype : a constructive generality

Genotype can be seen as a set of instructions carried within an organism's genetic code, the DNA. This is straight-forward. Phenotype's definition however is an example of constructive generality. According to the Free Dictionary, phenotype is "*the set of observable characteristics of an individual (e.g. cell) resulting from the interaction of its genotype with the environment*".

The question is at what temporal and spatial scales such a characteristic is observable, hence, measurable? Furthermore, should the measurements be integrated over time and cell population size and at what differential amount (resolution)?

Many cell processes are oscillatory. The frequency and amplitude of these oscillations as well as other qualitative characteristics, may also be considered as phenotype. When cells are studied individually or in sub-populations, these oscillations are observed. However, when a population of cells is studied as a whole, via, say a Western blot, it is quite possible that the observed result of these processes averages (in contrast to the median) to some behavior that does not exist at all anywhere in the population (Batchelor, Loewer & Lahav 2009). It is also possible that for some of these oscillatory processes and depending on cell (a)synchronicity, the observed population average is constant over time whereas the signal from individual cells or sub-populations is variable (Spiller, Wood, Rand & White 2010). On the other hand cells rarely do exist in isolation or alone and thus understanding interactions of cells at individual and population level is critical, particularly in areas such as cancer and tumor biology.

In section 0.7, it is proposed that what it is perhaps a more suitable assessment

of cellular phenotypes is pertaining changes to phenotype rather than absolute phenotypes; defining cell behavior as a trajectory through phenotype states or ensembles of states may be more useful for biological systems.

0.3 Cellular noise

The question why cells with the same genes/genome while exposed to the same environment do not always exhibit the same phenotype is worth considering. The answer to this is most likely not found in further end-point genetic sequencing, though potentially more insight into genome dynamics would be useful. Rather it is the combination of noise and the diversity in behavior of otherwise deterministic dynamical systems; not an anomaly but a fact of life.

Unfortunately the prevailing deterministic thinking in cell biology, as (Quaranta & Garbett 2010) put it, has obscured the importance of variability and noise in the cell's behaviour and micro-environment by averaging out the phenotypes of large populations of cells. In fact, for many tissues the variation in gene and protein expression among clonal progenitor cells has biological consequences as it was shown that it leads to different functional states (Chang, Hemberg, Barahona, Ingber & Huang 2008).

The fluctuations in the expression of a gene arise mainly from two sources of stochasticity. Extrinsic sources are external to the cell and due to existing cellular heterogeneity, e.g. cell size, cycle stage, number of ribosomes and growth rate.

On the other hand, intrinsic sources are internal to the cell. They have to do with the stochasticity of chemical reactions at the molecular level particularly when low copy numbers of reactants are involved. Therefore, chemical reaction events leading to protein expression, such as translation and transcription across different but genomically identical cells will fluctuate even if the cells are otherwise in (cell cycle) synchronicity (Swain, Elowitz & Siggia 2002).

In (Elowitz, Levine, Siggia & Swain 2002), Elowitz et al studied the variation in expression of two genes (Cyan Fluorescent Protein, CFP, and Yellow Fluorescent Protein, YFP) within single cells of E. coli as the amount of intrinsic noise increased. It is important to note how they controlled the levels of intrinsic noise using internal reporter gene construct.

They observed that when intrinsic noise is low, the levels of CFP and YFP vary over time but in complete step within cells causing cells to exhibit a homogeneous color (the synthesis of CFP and YFP frequencies).

Increasing intrinsic noise caused the protein levels to vary independently within cells which resulted in different color depending on the expression of which CFP or YFP was at that particular time higher.

Cell shape characteristics such as nucleus size and perimeter shape do affect the levels of intrinsic noise in the cell but more importantly, it can be argued that they can localize or isolate it to certain cell areas. Cell shape determines

the physical characteristics of the volume within which the chemical reactions take place and thus controls the level of intrinsic noise because some areas will be more restricted and crowded than others. This also relates to the solid state chemical nature of many sub-cellular molecular compartments.

0.3.1 Maximum information flow and noise

In (Tkačik, Walczak & Bialek 2009), the problem of information flow in genetic control circuits is explored. A simplified version of these circuits consists of a transcription factor binding to DNA, regulating the transcription of mRNA and the further synthesis of one or more proteins, which are assumed to not interact with each other. Information is represented by the various molecular events flowing from the input (the transcription factor concentration) to the outputs (the synthesis of the protein molecules). In this model, there are two noise components; the input noise is due to fluctuations in the arrival and concentration of the transcription factor and the output noise which is proportional to the mean levels of expressed protein.

The authors assume (among other things) that the cell optimizes these processes for maximum information transmission (Tkačik, Callan & Bialek 2008) at a given maximum input concentration by the adaptability of the process to its input. The suggested framework is to maximize the mutual information between the input and output quantities constrained by the maximum number of molecules at both ends. The expression for the mutual information involves two probability distributions; the (input) transcription factor concentration; and the conditional probability of protein expression levels (outputs) given an input concentration. The overall distribution of protein expression levels is the product of the aforementioned distributions integrated over all input concentrations.

0.4 Genome evolution, protein families and Phenotype

It has been observed that different proteins, even across species, share 'similar' amino-acid sections. Such similarities indicate that the species the *homologue* proteins belong to, although now genetically isolated, had in the past shared a common ancestor species or that through a past gene duplication identical proteins in the same species became distinct but 'similar' through accumulated mutations.

Although proteins can be described in at least three levels; 3D structure, sequence structure and function, the main tool for family identification is structure-based sequence alignment as this is, from a computational point of view, the most tractable.

Exactly because of this multiple character of proteins, the notions of similarity / distance and relationships between protein domains or between protein linear motifs become difficult to define. The field of Information Theory has

provided a variety of tools and methods for quantifying protein similarity, detecting evolutionary conservation and mapping protein sequence structure to protein function. For example, the concept of mutual information used in Statistical Coupling Analysis (SCA). In (Socolich, Lockless, Russ, Lee, Gardner & Ranganathan 2005) it was shown that it is not safe to assume that two residues occupying their respective positions in the structure of a protein are statistically independent events.

SCA was used to identify pair-wise dependencies between sites in proteins of the same family by analyzing their statistical profile. The probability that a pair of sites has mutually co-evolved increases as its mutual information, essentially a correlation metric, increases. This implies that protein structural stability (folding) is undermined if a mutation in one of the sites is not compatible with its counterpart – even if they are not near in linear sequence space. Evidence of this coupling has been known for some time (Yanofsky, Horn & Thorpe 1964). However, the importance of this research lies in the suggested computational frameworks for detecting and quantifying protein site dependencies as well as sampling artificial proteins from a probability distribution using Monte Carlo and the Metropolis algorithm. The benefit of using the Monte Carlo method is to be able to *sample* the space of eligible proteins taking into account co-evolution dependencies without constructing the protein model as an actual rule-based system.

Yet, this is just the tip of the iceberg; the heart of the problem is to be able to determine the relationship between form (sequence) and function (impact on phenotype, which in turn most likely depends on the interaction network of the protein at the time of activation / regulation) of proteins emerging from the vast number of combinations of amino-acid interactions and in view of the multitude of different sequences associated with similar function. We argue that the form and function dyad, epitomized by genotype and phenotype, is a recurring theme in biology manifested at many levels as a dynamical system (see section 0.4.1) whose evolution (function) can potentially span a huge space but in reality, it is restricted to a relatively smaller number of attractor states, as its free parameters (form) change (see also section 0.5.2). In our opinion, the role of Information Theory and Statistical Mechanics is and will be significant in breakthroughs in cellular biology, as outlined below.

The aim in (Bialek & Ranganathan 2007) is to "*have a description of this ensemble of sequences, ideally being able to write down the probability distribution out of which functional sequences are drawn*". To this end, a link between Monte Carlo sampling of artificial proteins and Maximum Entropy (ME) models is made. A ME distribution is one which is as random as possible while, at the same time, consistent with prior information, i.e. experimental observations. In (Mora & Bialek 2010) an attempt is made to lay the problem within a Statistical Mechanics (SM) framework because "*maximum entropy models naturally map onto known statistical physics models, which will ease the study of their critical properties*".

This direction has good potential because it allows a whole class of *emergent* systems as diverse as networks of neurons, ensembles of (amino-acid) sequences and flocks of birds to be discussed within a single, unified framework with powerful mathematical analysis tools which combine the key concepts of Information, Entropy and Energy. Moreover, this framework allows for the investigation of scale-free properties (see section 0.5.1) and self-organized criticality (see section 0.6.4) in these systems as demonstrated by Mora & Bialek.

0.4.1 Phenotype State Spaces

Using the example of the routes a cell may take during differentiation, Waddington (Goldberg, Allis & Bernstein 2007, Wang, Zhang, Xu & Wang 2011) devised the analogy of a landscape on which a ball under the force caused by some gravitational field, rolls down from high altitude, follows various (forking) paths with occasionally a brief upward direction given enough kinetic energy. Eventually the ball settles (or is trapped) in a low-energy valley and stays there until there is some change in the landscape which will get it rolling again.

Traditionally, Waddington's epigenetic landscape is employed as a demonstration of the variability but overall finiteness, discreteness and stability of phenotypes. Similarly, more complex phenotypes such as cell morphology or cell fate (proliferation/apoptosis) processes and development and differentiation may also be considered as operating in a space of phenotypic attractor states. A big challenge in modern biology is to link these to molecular implementations to predict biological systems behavior.

(Balázsi, van Oudenaarden & Collins 2011) proposes some revisions to Waddington's picture in order to properly describe cellular dynamics and reflect recent research:

- The concentrations of all molecules in the cell and all relevant environmental factors must add one dimension each to the landscape, rendering it super high-dimensional.
- The landscape is not rigid; each point is subjected to molecular noise of various magnitude and properties. In addition, cells or mutations to the genome may reshape the landscape due to cell-to-cell interactions and growth rate dependence of protein concentrations.

In addition to Balázsi's extensions we would argue that a critical parameter is the activation state of a molecule, e.g. a kinase which is crucial to relate its activity to networks and phenotype.

The idea of a landscape, a surface which represents a potential / fitness / energy / probability / etc. spanning the space encompassing all the possible factors that affect it, has been around for some time. Sewall Wright introduced such landscapes to Biology (Wright 1932).

In mathematical optimization one seeks to minimize or maximize a fitness function subject to the constraints and inter-dependence of its various input

parameters. The fitness function is represented as a $(N+1)$-dimensional surface over its N inputs. Its minimisation is the process by which the deepest valley – the global minimum is sought. Naturally, one assumes the existence of a fitness function in the first place. This means that the relationship of all inputs is known and can be expressed analytically as a mathematical equation or at least be consistently computable.

Given that the possible search space is potentially extremely large – depending on the number of input dimensions and complexity of the fitness function – one can not rely on *global* algorithms which require enumerating and assessing each possible combination of inputs. In fact, the phrase "each possible combination" has little meaning when we are talking about continuous-valued inputs. Discretisation of input parameters may be of some aid but the success is problem dependent as the surface can be extremely multi-modal in which case solutions will be missed.

Various *local* search algorithms have been proposed for finding the deepest valley in such landscapes. A widely used one is "*gradient descent*" or "*hill climbing*" depending whether we seek the deepest valley (e.g. minimum energy) or the highest peak (e.g. maximum fitness). They are based on following the steepest way (as in steepest slope, gradient) up or down - in a sense it is the route water would follow, under the influence of gravity, when released from the top of a mountain.

A common characteristic of local algorithms is that they suffer from the problem of "*local minima*". Following the steepest path downhill guarantees that the next step will be at the same or lower altitude (fitness) than the current position. It does not guarantee however that the final step will be the lowest over the landscape. Simply because these algorithms are local. They are blind to the wider landscape topography because in order to see it, they will have to calculate it and it is too expensive to compute and store the landscape in more locations other than those on a simple trajectory.

Heuristics are workarounds to the inherent limitations of these (local) search algorithms. They are what we call *rules of thumb* and they may work most of the time but not all the time. So there are numerous heuristics to cope with the problem of *local minima* mostly based on the observation that optimization is not necessarily a monotonic process, i.e. one has to go down first in order to go further up (e.g. snakes and ladders).

But first consider an analogous problem from metallurgy (there are similar examples in spin-glass theory). Crystal formations in a metal at room temperature are fixed. Many of the properties of the material depend on bond stability which depends on the type of the crystal formations and their potential energy. The lower the energy the more stable the crystal structure. When the material is at high temperature, basically melting, the molecules are free to move around, break old and create new crystal formations of various types. As the temperature is reduced, molecular mobility is less and crystal structures become more permanent bar some local movement - the material slowly solidifies

Molecules are trying to solve a problem; the objective is to minimize the sum of the potential energy of the crystal bonds in the material given a set of N parameters: molecule positions, etc. These are the N dimensions over which the energy landscape lies. The collection of all molecules, the system, seeks the deepest valley of this landscape.

Annealing is a heuristic which improves this process by carefully controlling the temperature which is related to the mobility of molecules which is related to the probability of accepting a relatively bad solution now as a means of ending to a better solution later. This is equivalent to shaking the energy landscape and causing the, familiar by now, ball to jump out from the valley it has currently settled in, with a good chance that when it goes up, it will find its way to an even deeper valley. As the temperature is reduced, the landscape shaking is reduced. It can not give the ball sufficient energy to make these dramatic bounces but it can be enough to move the ball to the left or right a bit in order to stabilize its position locally.

Simulated Annealing in mathematical optimization, is a similar process by which the landscape is explored in an "*educated*" yet stochastic way. Given a current state, S_0, a candidate state, S_c, and a temperature value, T_t, the probability of moving to the S_c depends on the energies of the two states, E_0 and E_c and the temperature. At higher temperatures, the likelihood of jumping to the candidate state is very high even if its energy / fitness is not as good as the current state. This allows for sampling the landscape uniformly for good solutions. As the temperature is reduced however, candidate states with worse fitness than the current one are not likely to be accepted. Because we are using a computer, at each step, the best solution found so far is saved. Although this is not possible in physical systems e.g. metallurgy, it may be possible in biological systems, in fact there seems to be evidence pointing in this direction in epigenetic studies of glucose metabolism in yeast.

In mathematical optimization, raising the temperature is equivalent to injecting noise into the system. It is like shaking a pinball machine; making the behavior of sampling more random, more explorative, forcing the ball to jump out of local minima traps and via one or more bounces through possibly worse positions, to land to, hopefully, a better solution.

The geometry of the landscape depends on the properties of the system under study – dimensionality, input relationships, etc. – and the success of local search algorithms crucially depends on geometry. Take for example the size of the basin of a valley. If deep valleys (good solutions) have a very narrow basin, then the probability that a ball bouncing randomly around its neighbourhood falls in it, is low; the solution will be missed. Another landscape feature is its "*ruggedness*" – at one extreme there is the Fuji-mountain type of landscape where all routes will eventually lead the ball to the best solution. At the other extreme there are the rugged landscapes, characteristic of discrete-input or combinatorial problems, e.g. protein sequence space. Rugged landscapes are characterised by a large number of neighbouring peaks and valleys and where a small step has an

enormous effect on fitness; thus making it difficult for local search algorithms to succeed. A measure of the "*ruggedness*" of a landscape is its *correlation length value* (Stadler & Schnabel 1992). This is a simple metric of the number and distribution of its peaks and valleys in relation to their neighbours. A series of fitness values, i.e. the height of the landscape, is collected by doing a random walk on it; sampling is necessary because it is not practically feasible to exhaustively enumerate the whole landscape. The *correlation length value* is the average correlation between two such values a number of steps, t, apart. As observed in (Hordijk & Kauffman 2005), "*smooth landscapes have long correlation lengths, random landscapes have zero correlation length, and rugged landscapes have correlation lengths that decrease as ruggedness increases*". See also 0.6.6

0.4.2 Evolution as optimization

Evolution is a way for living organisms to cope with environmental changes and continue living. The key elements are two.

Primarily, a source of stochasticity, noise, which will perturb the current state in order to explore the landscape of solutions as efficiently as possible.

Secondly, some modes of exploration rely on the existence of memory (i.e. genetic blueprint, DNA or epigenetic information or even biochemical attractors) in order to accumulate fitness and / or recombine fitness parameters with mates thus making this exploration wider. The fitness parameters in genetic memory may be passed on to the following generations or be erased when the organism or cell dies.

How the exploration of the space is happening and whether memory is used or not, it all depends on the life span and scale of an organism and the afforded energy to be expended on finding a fit solution.

The need for survival in a changing environment is the same for animals as is for single cells. Each in its own environment will seek to adapt. Cellular intrinsic noise aids robustness and stability within the cell in the face of micro- and macro-environmental changes in the same way as noise in the analogy of the ball rolling on a landscape, aids the search for optimality in the vast and complex space of candidate solutions. Phenotypic variation is key to the survival of cells, as is the fact that these phenotypes are short-lived. Noise levels influence the exploration of phenotypes in three interconnecting ways: i) energy; how much of it is expended in unfit candidates, ii) time to react / adapt to environmental changes, iii) variability of phenotypes; how "adventurous" is the exploration.

0.5 Complex Networks

One approach to modeling interactions between several players be they species in an evolutionary context or molecules in a biological processes context are the so-called Complex Networks. Using Graph Theory terminology, Paul Erdös and

Alfréd Rényi (Bollobas 2001) derived a framework for analysis of networks with random connectivity – what they called Random Graphs. The emphasis here is on qualitative aspects of the connectivity of interactions between players, e.g. degree, clustering coefficient etc., rather than how one player interacts with the other (e.g. the Cellular Automata rules, above, or Random Boolean Networks, below). One such aspect is the probability that a node has a certain number of links – the degree. The degree distribution is a key feature in categorizing networks. For example, in a plain *random regular* network each node has a constant degree, whereas Erdös-Rényi have normal (Gaussian) degree distribution.

0.5.1 Power law forms, small-world and scale-free networks

Scale-free networks (Barabasi & Albert 1999) are those whose degree follows a *power law* distribution; the probability of degree k, $P(k)$, is exponentially decreasing with degree, $P(k) = Ak^{-\gamma}$, $\gamma > 0$ is a constant that determines the shape of the distribution; A is a normalizing constant. This means that, depending on γ, a large fraction of nodes have a small number of links and only a few – the so called hubs – have significantly higher, by orders of magnitude, connectivity. Popularization of the term has certainly benefited from the fact that many random, complex networks occurring in our every day life, e.g. social interaction networks or the world-wide web itself fall in this category. Nonetheless, networks whose fundamental construction properties, e.g. node degree, are drawn from power law distributions, are significant because their structure remains (statistically) unaltered when these construction properties change, even by orders of magnitude. This stems from the unique property of power law distributions that $P(nk) = A(nk)^{-\gamma} = n^{-\gamma}Ak^{-\gamma} = n^{-\gamma}P(k)$ which is a scaled version, by $n^{-\gamma}$, of the original distribution, $P(k)$. Hence the term *scale-free* networks.

On the other hand, networks are also characterised by the average shortest path between any two nodes; the least number of intermediate nodes connecting them. This property increases modularity and affects the speed of signal propagation as well as signal-to-noise ratios. A small-world network is one with small average shortest path values (typically proportional to the logarithm of the total number of nodes) and is usually characterised by abundance of sub-networks with high intra-connectivity and low inter-connectivity, mainly via the hubs.

0.5.2 Controllability of Complex Networks

One interesting question about Complex Networks is controllability. Given a current state can the network be steered towards a desired state, by which nodes and how? To this end, Barabasi and colleagues (Liu, Slotine & Barabasi 2011) have used elements of control theory and statistical physics to analyse the controllability of several types of Complex Networks. A practical finding of their work is that *driver nodes* (those nodes which can control the whole system) in the networks investigated tend to avoid the hubs and are mostly found among the

ranks of nodes with low degree. In general, this is an interesting and somewhat counter-intuitive finding laying a general framework for such analysis. However, its applicability to biological networks, e.g. transcriptional networks controlling cellular phenotypes, has been disputed recently by (Muller & Schuppert 2011) who argued that these particular types of networks show a high degree of co-regulation, i.e. when regulators partner with each other, as a linear combination, to control common targets, see also (Bhardwaj, Carson, Abyzov, Yan, Lu & Gerstein 2010). The result is that the actual gene expression state space is reduced to a combinatorial expression space with relatively low dimensionality, hence controllability can be achieved by only a few nodes.

0.6 Random Boolean Networks

0.6.1 Cellular Automata

Boolean variables are those which are permitted only two states, usually denoted as 1 / 0 or, more intuitively, ON / OFF. Boolean functions are functions in terms of boolean variables and the three boolean operations (conjunction, disjunction and negation a.k.a. AND, OR, NOT). A boolean logic element is a computational unit of one or more inputs, one output and a boolean transfer function which maps the inputs to the output. In simple terms, a boolean element may be ON or OFF depending on what the state of its inputs are. Once its output state is determined, it will affect all those elements which is connected to turning them ON or OFF respectively.

Early computational models of evolution were based on Cellular Automata (CA). These consist of a set of boolean logic elements arranged on a regular grid. The behavior (output state) of each element is determined by the behavior of its immediate neighbours via simple rules, for example "*turn ON if exactly 2 or 3 of your neighbours are ON; otherwise turn OFF*".

Random Boolean Networks (RBN) are an extension to CA where the behaviour of each node in the network is no longer determined by only its immediate neighbours but by potentially any other node irrespective of grid-distance. In a sense, in RBN there is no grid. CA and RBN are useful in studying bottom-up, emergent behavior stemming from simple elements operating under simple rules. However, emulating emergence in nature using these techniques may be too simplistic and fundamentally constrained by implementation in digital computers.

0.6.2 NK automata

Cellular chemical chain reactions are often explained using nomenclature more appropriate for static relationships between chemical reactants and their product under normal laboratory conditions, for example often they are wrongly called cascades or pathways (Jørgensen & Linding 2010).

In reality, chemical reactions within the cell are highly dynamical and non-linear, they influence each other via cross-talk and occur in conditions far from ideal (laboratory); any attempt to describe them with static, deterministic terminology will lead to over-simplification.

Cell states and the transition between them has also received the same treatment of being characterised by over-simplistic, top-down models where in fact they are attractors of bottom-up, stochastic processes with complex emergent properties.

One of the first attempts to put them in a more "appropriate" framework was done by Kauffman (Kauffman 1969, Kauffman 1993) who used random boolean networks to model cell state and regulatory interactions in terms of epigenetic factors, genes and proteins.

Kauffam's NK automaton consists of N boolean logic elements with K inbound connections. The state of an NK automaton at a given time is basically the state of each of its elements. Such a system is dynamic in the sense that it contains feedback loops. When the output of an element changes in response to its inputs changing, it will spawn a number of changes in other elements which it connects to, These changes in turn will cause the inputs to the first element to change and so on. Determining the stability of such a system is tricky.

By analogy, cell state or phenotype is determined by the state of the individual elements (epigenetic factors, genes or proteins). However, in the case of the cell, a state, say proliferation or cell-cycle check points, may correspond to a number of actual molecular states because the available means of observation and poor resolution limit our ability to differentiate between cell states.

The NK automaton serves less as a cell modeling paradigm and more as a conceptual framework where the behavior and, in particular, criticality of a dynamical distributed system is investigated. A NK automaton has 2^N possible states – the enumeration of all possible combinations of its elements being ON or OFF. This number is growing exponentially with the number of elements; a relatively small N yields a prohibitively large number of possible states. However, a lot of possible states may never be reached.

The free parameters of the NK automaton are N, K and the boolean functions of each element (e.g. the "rules"). We tend to study the system through aggregating a fairly large number of simulations and not on a one-off basis. E.g. we try to draw conclusions about the long-term average behavior of a system when the complexity of the topology, or the complexity of the rules is of a certain degree and drawn from a certain distribution. The initial conditions (i.e. starting state) are important in determining which cycle the system will settle in, only in cases when there are more than one cycles. Because these systems are completely deterministic, once inside a cycle, the starting state is completely unimportant. For example, if state A leads to B which leads to C, then starting either from A or B, the end state will still be C. Which also means that it does not matter via which state the system enters a cycle should there be more than one entry states.

The most important parameter is K – this is the number of other elements an element is influenced by. Effectively, it controls the complexity of the system by influencing the number of cycles and cycle lengths. Whereas the parameter N lays down the players by controlling the total number of possible states, K effectively controls the complexity of the system by arranging these 2^N possible states in loops and attractors. However, as N becomes large the way to find these attractors is usually by random sampling which gives a statistical indication of their existence. For example, for a long time it was wrongly thought that for the NK automata of $K = 2$, the number of attractors was proportional to $\sqrt{N}$ (Socolar & Kauffman 2003). Just because a specific attractor is difficult to discover by random sampling (e.g. it is small) it does not mean that it can not be manifested. This last point is important for coupled biological dynamical systems (as in cells in the same tissue) which through evolution have adapted to probable attractor behavior from their peers and when a rare yet perfectly possible attractor appears the system splinters.

At this point it is useful to introduce another aspect of complexity beyond the number of cycles and cycle lengths. This is the similarity between states, more importantly between states of the same cycle. How dramatic are the state transitions within a cycle? There are many distance metrics which output the distance between two points in a space, for example the Euclidean metric. A metric which is appropriate for boolean (or discrete-valued in general) variables and spaces is the Hamming distance. It is defined as the number of bits the two states differ by. Other metrics do exist and can be used instead.

A value of $K = 1$ yields an un-interesting unit-length cycle, meaning that the system settles to a single state. In biological systems this is the equivalent of death – nothing changes. On the other hand, large values of K yield extremely long cycles through a very large number of states. The system is very complex, almost pseudo-random although it is totally deterministic – this is a common feature of Chaotic systems. K values of 2, 3, 4 yield a small number of cycles and cycle lengths.

There are no magic numbers, this is just what makes sense to the human brain and its own pattern recognition capacity, given the temporal and spatial resolution of observation and measurement. Extracting general rules from the behavior of these networks may be more useful. It is also time to rethink how a cellular phenotype corresponds to the behavior of such automata. Is a phenotype mapped to a certain state occurring at the i^{th} step of the cycle? Or is phenotype a subset of cycle states or indeed the whole cycle itself? After all a phenotype has an effect inside and outside the cell. It is useful to study these systems interacting. If there was ever a resemblance between a Random Boolean network, an NK automaton, a cellular automaton etc. with the mechanism inside a cell, then it is worth investigating what happens when a lot of these systems are coupled as in interacting cells within the same tissue. The objective is for each unit alone and the ensemble as a whole to be stable yet to exhibit rich and diverse behavior which will allow for it to adapt to environmental changes, to share resources

and in general to sustain existence not only in single units but as an ensemble. The question is what are the requirements in terms of their complexity – i.e. the number of cycles and cycle lengths? What is an optimum level of interaction between units?

0.6.3 Computation at the edge of chaos

The relationship between complexity and synthesis (as in Self-Organisation) was studied by John Von Neumann and Stanislaw Ulam in the 1940's and 50's in the context of his proposed self-reproducing automata (before the discovery of the DNA!). He observed that there exists a critical threshold of complexity below which the process of synthesis is degenerative but above which, synthesis may become "explosive". But there is another issue here which is as important as complexity and this is stability and criticality. A small perturbation to a dynamical system can have a temporary effect which soon disappears as the system returns back to stability or can have a substantial effect and cause the system to become unstable. There are one or more parameters of a dynamical system which control its behavior to small perturbations. For some critical value of these properties, the system's response to a small perturbation switches from stable to chaotic.

Chris Langton in his paper "*Computation at the edge of chaos*" (Langton 1990) proposed that Von Neumann's threshold is more likely to be a region; too little or too much complexity can kill synthesis. Langton made a distinction between two kind of cellular automata; those that exhibit a very ordered, predictable behavior and soon die or cycle between a few states (stable); and those that exhibit a totally unpredictable behavior (chaotic). He introduced a parameter (λ) as the fraction of rules which produce a "live" automaton over the total number of rules and argued that this parameter controlled the criticality of the generated automata. Some range of λ values produced really interesting automata with rich and diverse behavior – clearly distinct from that at the ordered regime but without becoming chaotic either. This region was called "*edge of chaos*"; what M.Waldrop calls "*the zone between stagnation and anarchy*" (Waldrop 1992).

On the point of system stability, it may seem counter-intuitive but strong stability will cause a system to be less sensitive to perturbations which effectively causes it to not react to environmental change. For this reason, behaviour at the *edge of chaos* is even more interesting.

One way to quantify complex behavior (as in output repertoire, e.g. in a morphological space) is by using Kolmogorov's notion of complexity. It is defined as the shortest description (program) which can replicate an output pattern exactly. Consider a system consisting of a uniform random generator which produced one million ones and zeros. According to Kolmogorov, its behavior is as complex as it gets because there is no program which can be used to produce the sequence exactly without resorting to just printing the sequence itself hard-coded in the program. Now consider a sorted version of this random sequence. A very sim-

ple program (print 500,000 zeros; print 500,000 ones) can be used to reproduce it, hence its complexity is minimal. Kolmogorov's notion of complexity is one of many metrics which can rank the complexity of the behavior of cellular automata and classify them as 'boring', 'interesting', 'chaotic' or anything between.

0.6.4 Self-Organized Criticality and Power Law Forms

The term *self-organized criticality* is associated with systems of locally interacting agents where occasionally a critical state is reached, there is a phase transition and global dynamics emerge. The example usually cited is that of creating a sandpile by dropping sand, one grain at a time, onto a surface, at random locations. At the beginning, each new grain settles in the vicinity of where it was dropped through local interactions. At this point, small, local avalanches can be observed. This continues until the slope of the pile reaches a critical value. This is when the drop of a single grain of sand will cause all sort of avalanches including global ones. The system changed by itself (self-organisation) from local to global dynamics (criticality). Furthermore, computer simulations of simplified sandpile models have revealed the distribution of avalanche magnitudes to generally follow power law forms (see section 0.5.1). Subsequently, power laws were (re-)discovered in other toy or real world systems, e.g. in the frequency and magnitude of earthquakes (Gutenberg-Richter), the Bak-Sneppen model of evolution, etc. (Bak 1996). In (Mora & Bialek 2010) a link is made between statistical mechanics and biological systems whose events exhibit power law distributions (see section 0.4)

0.6.5 NK automata variations

These automata systems are programmed for and run on digital, discreet time computers which contain a small number of processors. This means that the system is not reacting instantaneously to any input changes. Instead the processor is updating each sequentially, one after the other in a queue, and synchronously in step with its clock. This may have significant effects to the behavior of the system. For example, it was observed that the synchronous update scheme introduces a large number of unstable attractors (Greil & Drossel 2005). Simulation with parallel computers may or may not be able to overcome this limitation depending on their architecture, but there is definitely a limitation due to the small number of processors. A more natural solution would be an analogue, parallel computer consisting of many thousand of very simple analogue computational units (basically analogue electronics circuits) communicating via wireless broadcast. Insight from biological neural circuits may be helpful.

Networks which contain a mixture of real number and boolean variables are probably a more realistic implementation as they may be used to introduce protein expression levels. Accordingly the rules must be such so as to accommodate both types of variables. Alternatively, a type of variable which is a cross between

boolean and real is called *fuzzy*. In the framework called fuzzy logic, qualitative terms like *low*, *medium* and *high* are mapped to real value ranges. Fuzzy calculus allows fuzzy transfer functions. Fuzzy logic has been used to model various signaling systems (Aldridge, Saez-Rodriguez, Muhlich, Sorger & Lauffenburger 2009, Morris, Saez-Rodriguez, Clarke, Sorger & Lauffenburger 2011).

Other additions include:

1. stochasticity : noise as described earlier - signals may be distorted by this noise or never arrive, rules may also be affected by noise, connectivity may be altered by noise.
2. probabilistic rules : the rules are probability distributions rather than deterministic functions.
3. Dynamic / context-based topology : connectivity between elements changes over time old links are broken, new links are created deterministically, over time or depending on the context (current state of the network, external influences)
4. (random) time delays in the propagation of signals : signals may take a while to arrive to their destination depending on current state or just randomly. They may never arrive.
5. the incorporation of some kind of memory : the current state may depend not only on current but also past states.

(Kadanoff, Coppersmith & Aldana 2002) contains a review of research in some of the above points.

0.6.6 Boolean networks and fitness landscapes

In (Hordijk & Kauffman 2005) a link is made between a fitness landscape, the players that affect it and also each other and the NK automaton. The free parameters of the landscape is the set of N genes which form a space into which the fitness landscape spans. Each gene combination – a genotype – is assigned a fitness value. This is represented by the height of the landscape at that particular point, which is the sum of individual gene contributions. Epistasis is modelled by the factor K which is the number of genes affecting a given gene.

The "*ruggedness*" of the landscape increases with K. When each gene is affected by all other ($K = N - 1$) genes, then the landscape is random with zero correlation length which means that moving from one genotype to another very near, the change in fitness is likely to be huge. Contrast this with small values of K which yield landscapes of long correlation lengths meaning that a small change in genotype will most likely have a proportionally small change, positive or negative, in fitness. Undoubtedly, the methods to explore the landscape and maximize fitness will differ according to "ruggedness".

As a means of modeling co-evolution among different species, (Kauffman 1995) suggests that genes from one species affect the fitness of another species via some kind of inter-species, external epistasis. This is, for example, the case where a prey

species develops a particular skin color as a defense mechanism for escaping its predator. In effect, the prey has deformed the fitness landscape of the predator. The predator is no longer as fit as it used to be because, say, it can not spot that color easily, from a distance.

The NKC model consists of a number of NK automata, one for each species. Each automaton has N genes (boolean elements) and each gene is affected by K other genes of the same species as before. However each gene is also affected by C external genes, from the other species.

The *edge of chaos* is a recurring theme in dynamical systems and appears also in coupled fitness landscapes. Depending on the parameters K and C, co-evolution may behave in the ordered or chaotic regimes or in between, at the edge of chaos. In the ordered regime all species reach some acceptable fitness level and happily co-exist without further increasing their fitness. This is happening when K is high or C is low. If K is high then a species landscape is extremely rugged so the effect of losing fitness because of the effect of other species (via C) can quickly be compensated as a new fitness peak is bound to be in the neighbourhood. Likewise, if C is low, then the influence of one species to another's fitness landscape is minimal, therefore if a species finds a fitness peak, it is difficult to lose it because of the effect of another species.

In the chaotic regime, the species are changing constantly and never settle down because they affect each other too much. The effect of a low K value is that the fitness landscape will be quite smooth with very few fitness peaks very far apart. Evolution for this species will be a long and slow process. If another species continuously deforms this landscape (via C) the long and slow process of evolution will never bear any results as it will continuously be rebooted, the players are expending more energy in "sabotaging" each other than in improving their own fitness. It is claimed that maximal fitness for all species involved lies in the region between order and chaos – the *edge of chaos*. This region can be reached by adjusting K and C values.

It can't be stressed enough that automata serve as a tool of investigating dynamical systems via simulation, rather than as faithful models of Nature. Let us not forget that gene regulation networks are the result of years of evolution and the manifestation of the laws of physics. This is in contrast with the random connectivity of RBN although there is research in using experimental data to form connectivity (Harris, Sawhill, Wuensche & Kauffman 2002).

0.7 Genomic State Spaces

The term state-space is often associated with dynamical systems and Ergodic Theory. This is the relationship between all the parameters (degrees of freedom) that determine the behavior of a system or process. In the classical example of a pendulum, the two parameters which determine fully its behavior are the angle from the vertical and angular velocity (which is the rate of change of the angle

with time). These two parameters make up the state-space of the system which can be visualized as a (two dimensional) plot of velocity versus angle. At the intersection of each possible combination of velocity and angle value, a mark is made. This will reveal the trajectories of the system or where does the system goes next given a current state. This is much more useful than a time plot because it gives a view of the system completely independent of initial conditions.

(Huang 2010) constructs a "genomic" state-space for a cell where the free parameters of the system are all the genes in the cell's genome. A gene expression pattern is a point in this vast space. Given no underlying gene regulatory interactions this space would be occupied totally by expression patterns – any expression pattern would be possible.

In reality there are several constraints in the interaction of genes, be they physical or chemical, what is called "gene regulation". The result of this is that only certain gene expression patterns are possible; only some states in the state-space can ever be occupied.

A trajectory in this state-space is a movement from one gene expression pattern to a neighbouring one and may constitute a phenotype change. This change is independent of time; there is no time variable in the state-space construction. All it is, is that it is possible to move from state A to state B if states A and B are connected via a trajectory. Whether this is statistically significant or not is another issue, related to the assumption of ergodicity.

An "attractor" in a state-space is a closed-loop trajectory which traps the state of the system in a cyclical behavior. For obvious reasons, a finite state-space must have at least one reachable attractor. The set of states which lead to the attractor, but not necessarily part of it, form the so-called "basin of attraction". In the case of a ball inside a bowl, the attractor is the single point at the bottom of the bowl. The basin of attraction is each point on the inside of the bowl.

It must be noted that a state-space is not the same as Waddington's epigenetic landscape. The main difference being that the latter has one extra dimension and this is the fitness or some other potential. That's why a state-space has trajectories, whereas a fitness landscape has hills, valleys, canals and ravines.

The problem with building a genomic state-space is that it requires gene expression data be obtained accurately for a single cell over a long time and at high temporal resolution. If the system is ergodic (see section 0.7.1), then the genomic profile of a lot of single cells at a specific time is equivalent to that of a single cell over a long time. This may be easier but practical issues still remain.

0.7.1 Ergodic theory

The question whether observations made on a population at a particular time instance as a snapshot are equivalent to observations made on a small subset of the population averaged over a long time interval always arises or must arise in experimental setups. In effect, one asks whether what has been observed at a

snapshot is not significantly different to the long term (over time) behavior of each member of the population.

The answer is of course, yes and no. Under some conditions, the observations are equivalent. Such a system is called ergodic. Ergodic Theory's main question is to find those conditions, if any, under which ergodicity holds.

Those points (states) in the state-space which are occupied by a unique trajectory are ergodic sets as they satisfy the main prerequisite which is that once in, the probability of leaving it is zero. An alternative, equivalent prerequisite is that each point (state) in an ergodic set must happen with equal frequency as any other.

In many biological experiments where some aspects of a phenotype corresponding to a given genotype are measured, it is necessary to question the relationship between time and space averages and thus ask whether observing a snapshot of a lot of cells, as in measuring the phenotype of thousand of cells in a static microscope picture is equivalent to following a few of these cells around and analyzing their phenotypes over time.

It is thus possible to gain insight into the behaviour of individual cells as estimated from the population average while systematically down-sampling the fraction of cells analyzed of the population with an ensemble technique (e.g. mass-spectrometry).

To this end, stochastic profiling is a method to quantify single-cell heterogeneities without the need to measure expression in individual cells (Janes, Wang, Holmberg, Cabral & Brugge 2010). Instead, measurements are made over a number of smaller samples of cells collected randomly from the large population. Heterogeneously expressed genes will have a larger variance when compared with homogeneously expressed genes acting as a reference (and identified by smaller variance). This is a better alternative to either observing the whole population which invariably averages out any single-cell heterogeneity or observing individual cells which entails large measurement errors.

Bibliography

Aldridge, B. B., Saez-Rodriguez, J., Muhlich, J. L., Sorger, P. K. & Lauffenburger, D. A. (2009), 'Fuzzy Logic Analysis of Kinase Pathway Crosstalk in TNF/EGF/Insulin-Induced Signaling', *PLoS Comput Biol* **5**(4), e1000340+.

Bak, P. (1996), *How Nature Works: The Science of Self-organized Criticality*, 1st edn, Copernicus (Springer).

Balázsi, G., van Oudenaarden, A. & Collins, J. J. (2011), 'Cellular Decision Making and Biological Noise: From Microbes to Mammals', *Cell* **144**(6), 910–925.

Barabasi, A. L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science (New York, N.Y.)* **286**(5439), 509–512.

Batchelor, E., Loewer, A. & Lahav, G. (2009), 'The ups and downs of p53: understanding protein dynamics in single cells', *Nature Reviews Cancer* **9**(5), 371–377.

Bhardwaj, N., Carson, M. B., Abyzov, A., Yan, K.-K., Lu, H. & Gerstein, M. B. (2010), 'Analysis of Combinatorial Regulation: Scaling of Partnerships between Regulators with the Number of Governed Targets', *PLoS Comput Biol* **6**(5), e1000755+.

Bialek, W. & Ranganathan, R. (2007), 'Rediscovering the power of pairwise interactions'.

Bollobas, B. (2001), *Random Graphs (Cambridge Studies in Advanced Mathematics)*, 2 edn, Cambridge University Press.

Bruggeman, F. J., Westerhoff, H. V., Hoek, J. B. & Kholodenko, B. N. (2002), 'Modular response analysis of cellular regulatory networks', *Journal of theoretical biology* **218**(4), 507–520.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. (2008), 'Transcriptome-wide noise controls lineage choice in mammalian progenitor cells', *Nature* **453**(7194), 544–547.

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002), 'Stochastic Gene Expression in a Single Cell', *Science* **297**(5584), 1183–1186.

Fisher, J. & Henzinger, T. A. (2007), 'Executable cell biology', *Nature biotechnology* **25**(11), 1239–1249.

Goldberg, A., Allis, C. & Bernstein, E. (2007), 'Epigenetics: A Landscape Takes Shape', *Cell* **128**(4), 635–638.

Greil, F. & Drossel, B. (2005), 'Dynamics of Critical Kauffman Networks under Asynchronous Stochastic Update', *Physical Review Letters* **95**(4), 048701+.

Harris, S. E., Sawhill, B. K., Wuensche, A. & Kauffman, S. (2002), 'A model of transcriptional regulatory networks based on biases in the observed regulation rules', *Complexity* **7**(4), 23–40.

Hordijk, W. & Kauffman, S. A. (2005), 'Correlation analysis of coupled fitness landscapes', *Complexity* **10**, 41–49.

Huang, S. (2010), 'Cell Lineage Determination in State Space: A Systems View Brings Flexibility to Dogmatic Canonical Rules', *PLoS Biol* **8**(5), e1000380+.

Janes, K. A., Albeck, J. G., Gaudet, S., Sorger, P. K., Lauffenburger, D. A. & Yaffe, M. B. (2005), 'A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis', *Science* **310**(5754), 1646–1653.

Janes, K. A., Wang, C.-C. C., Holmberg, K. J., Cabral, K. & Brugge, J. S. (2010), 'Identifying single-cell molecular programs by stochastic profiling.', *Nature methods* **7**(4), 311–317.

Jørgensen, C. & Linding, R. (2010), 'Simplistic pathways or complex networks?', *Current Opinion in Genetics & Development* **20**(1), 15 – 22. Genetic and cellular mechanisms of oncogenesis.

Kadanoff, L., Coppersmith, S. & Aldana, M. (2002), 'Boolean Dynamics with Random Couplings', *ArXiv Nonlinear Sciences e-prints* .

Kauffman, S. (1995), *At Home in the Universe - The Search for the Laws of Self-Organization and Complexity*.

Kauffman, S. A. (1969), 'Metabolic stability and epigenesis in randomly constructed genetic nets', *Journal of Theoretical Biology* **22**(3), 437–467.

Kauffman, S. A. (1993), *The Origins of Order: Self-Organization and Selection in Evolution*, 1 edn, Oxford University Press, USA.

Langton, C. (1990), 'Computation at the edge of chaos: Phase transitions and emergent computation', *Physica D: Nonlinear Phenomena* **42**(1-3), 12–37.

Linding, R. (2010), 'Multivariate signal integration', *Nature Reviews Molecular Cell Biology* **11**(6), 391–391.

Liu, Y.-Y., Slotine, J.-J. & Barabasi, A.-L. (2011), 'Controllability of complex networks', *Nature* **473**(7346), 167–173.

Mora, T. & Bialek, W. (2010), 'Are biological systems poised at criticality?'.

Morris, M. K., Saez-Rodriguez, J., Clarke, D. C., Sorger, P. K. & Lauffenburger, D. A. (2011), 'Training Signaling Pathway Maps to Biochemical Data with Constrained Fuzzy Logic: Quantitative Analysis of Liver Cell Responses to Inflammatory Stimuli', *PLoS Comput Biol* **7**(3), e1001099+.

Muller, F.-J. & Schuppert, A. (2011), 'Few inputs can reprogram biological networks', *Nature* **478**(7369), E4.

Quaranta, V. & Garbett, S. P. (2010), 'Not all noise is waste.', *Nature methods* **7**(4), 269–272.

Socolar, J. E. S. & Kauffman, S. A. (2003), 'Scaling in Ordered and Critical Random Boolean Networks', *Physical Review Letters* **90**(6).

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005), 'Evolutionary information for specifying a protein fold', *Nature* **437**(7058), 512–518.

Spiller, D. G., Wood, C. D., Rand, D. A. & White, M. R. H. (2010), 'Measurement of single-cell dynamics', *Nature* **465**(7299), 736–745.

Stadler, P. F. & Schnabel, W. (1992), 'The landscape of the traveling salesman problem', *Physics Letters A* **161**, 337–344.

Swain, P. S., Elowitz, M. B. & Siggia, E. D. (2002), 'Intrinsic and extrinsic contributions to stochasticity in gene expression', *Proceedings of the National Academy of Sciences* **99**(20), 12795–12800.

Tkačik, G., Callan, C. G. & Bialek, W. (2008), 'Information flow and optimization in transcriptional regulation', *Proceedings of the National Academy of Sciences* **105**(34), 12265–12270.

Tkačik, G., Walczak, A. M. & Bialek, W. (2009), 'Optimizing information flow in small genetic networks'.

Waldrop, M. M. (1992), *Complexity: The Emerging Science at the Edge of Order and Chaos*, Simon & Schuster.

Wang, J., Zhang, K., Xu, L. & Wang, E. (2011), 'Quantifying the Waddington landscape and biological paths for development and differentiation', *Proceedings of the National Academy of Sciences* .

Wright, S. (1932), 'The roles of mutation, inbreeding, crossbreeding and selection in evolution', *Proceedings of the Sixth International Congress of Genetics* **1**, 356–66.

Yanofsky, C., Horn, V. & Thorpe, D. (1964), 'Protein Structure Relationships Revealed by Mutational Analysis', *Science* **146**(3651), 1593–1594.

Zamir, E. & Bastiaens, P. I. (2008), 'Reverse engineering intracellular biochemical networks', *Nature chemical biology* **4**(11), 643–647.