

Chapter 2

Generic Properties of Dynamical Models of Protocells

2.1 Introduction

Models are of great importance for protocell research, not only for the usual reasons why models matter, but also because real protocells are not yet available in the lab. There are indeed some cases where one or a few duplications have been achieved (Hanczyc and Szostak 2004; Luisi et al. 2004; Luisi 2006; Stano et al. 2006; Schrum et al. 2010; Stano and Luisi 2010a) but so far, to the best of our knowledge, a sustained growth of a population of protocells has never been observed.

We will be particularly interested in models that allow us to explore the generic properties of protocells and of protocell populations. Of course, it is perfectly legitimate to concentrate on a particular hypothesis and to develop specific models well-suited to study its properties. But at the present stage of our knowledge we believe it can be even more important to be able to grasp the generic properties of these systems.

Protocells lie somewhere in between chemistry and biology: their ingredients are chemicals, as well as those of living beings. And their wished-for properties are indeed typical of life. That's why we find it appropriate to discuss here some features of models of biological systems aimed at describing some of their generic properties—a field of research that has been properly referred to as “complex systems biology” (Kaneko 2006).

Although it is widely agreed that “biological systems are complex”, there are several important features of the science of complex systems that have not yet deeply affected the study of biological organisms and processes. Indeed, biology has been largely dominated by a gene-centric view in the last decades, and the one gene—one trait approach, which has sometimes proved to be effective, has been extended to cover even complex traits. This simplifying view has been appropriately criticized, and the movement called systems biology (Noble 2006) has taken off. Systems biology emphasizes the presence of several feedback loops in biological systems, which severely limit the range of validity of explanations based

upon linear causal chains (e.g. gene→behaviour). Mathematical modelling is one of the favourite tools of systems biologists to analyse the possible effects of interacting negative and positive feedback loops which can be observed at several levels (from molecules to organelles, cells, tissues, organs, organisms, ecosystems).

Systems biology is mainly concerned with the description of specific biological items, like for example specific organisms, or specific organs in a class of animals, or specific genetic-metabolic circuits. Therefore, despite its usefulness in stressing the need for a systems approach, its focus is not concentrated on the search for general principles of biological organization, which apply to all living beings or to at least to broad classes.

We know indeed that there are some principles of this kind, biological evolution being the most famous one. The theory of cellular organization also qualifies as a general principle. But the main focus of biological research has been the study of specific cases, with some reluctance to accept (and perhaps a limited interest for) broad generalizations. This may however change, and it is indeed the challenge of complex systems biology: looking for general principles in biological systems, in the spirit of complex systems science that searches for similar features and behaviours in various kinds of systems. When speaking of protocells, one might perhaps prefer the term complex systems chemistry, but what really matters is the quest for general (or at least broad) principles, and simplified models may be a royal road to uncover such principles.

The actual working of some principles of this kind in real biological systems may be inferred from observations, and in Sect. 2.2 some data confirming this claim will be reviewed.

In order to explore new general ideas and models concerning the way in which biological systems work, an effective strategy is that of introducing simplified models¹ and of looking for their generic properties. This can be done by using statistical ensembles of systems, where each member can be different from another (although they all share some common properties), and by looking for those properties that are widespread. This approach, inspired by physics, was introduced many years ago in modelling gene regulatory networks (Kauffman 1969 but see Kauffman 1993, 1995 for a comprehensive discussion). Some important concepts and models of such generic properties will be described in Sect. 2.3.

Since the data and models of Sects. 2.2 and 2.3 provide evidence in favour of the existence and importance of generic properties, we will focus in the following Chap. 3 on how these concepts might be important for protocells, and we will show that the complex systems approach to these systems can be particularly interesting, providing useful stimuli to the experimenters. Before doing so, the last Sect. 2.4 of this chapter will summarize the main known facts about protocells that need to be taken into account in the development of generic models.

¹Like the pioneering chemoton model, described in Gánti (1997).

2.2 Generic Properties of Biological Systems: Data

Biologists have been largely concerned with the analysis of specific organisms, and the search for general principles has in a sense lagged behind. This makes sense, since generalizations are hard in biology, however there are also important examples of generic properties (in the sense defined in Sect. 2.1) of biological systems. Here we will briefly mention only two properties of this kind, namely power-law distributions and scaling laws, which can be observed by analysing existing data.

Power-law distributions are widespread in biology: for example, the distribution of the activation levels of the genes in a cell belongs to this class (see Kaneko 2006 and further references quoted therein). This means that the frequency of occurrence of genes with activation level x , let's call it $p(x)$, is proportional to x^{-g} where g is a constant positive exponent (see Fig. 2.1). Similar laws are found for other important properties, like the abundance of various chemicals in a cell. As it is well-known, power-law distributions differ from the more familiar Gaussian distributions in many respects, the most relevant one being a higher frequency of occurrence of

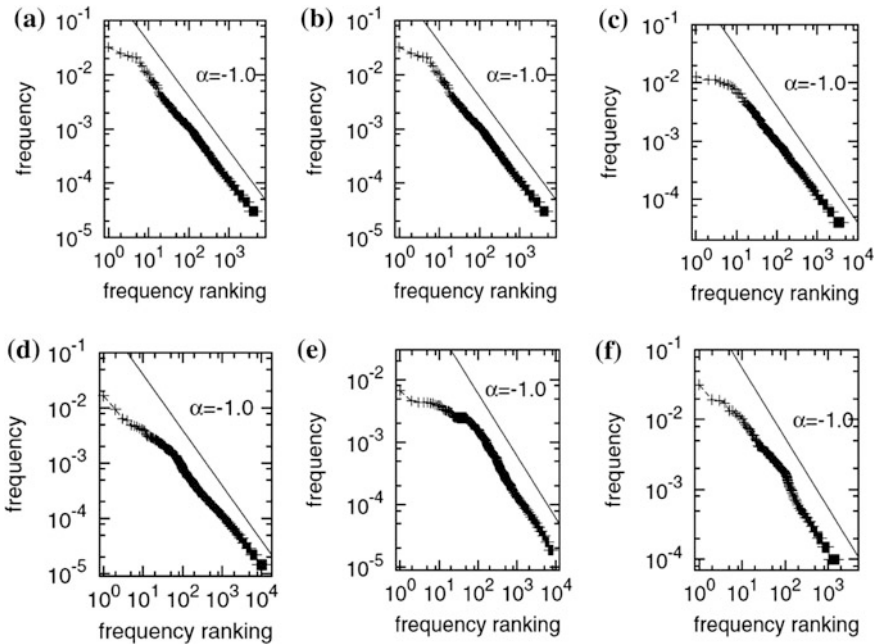


Fig. 2.1 Rank-ordered frequency distributions of expressed genes. **a** Human liver, **b** kidney, **c** human colorectal cancer, **d** mouse embryonic stem cells, **e** *C. Elegans*, and **f** yeast (*S. Cerevisiae*). The exponent of the power law is in the range from -1 to -0.86 for all the samples inspected, except for two plant data (seedlings of *Arabidopsis* and the trunk of *Pinus taeda*), whose exponents are approximately -0.63 . Reprinted with permission from (Furusawa and Kaneko 2003)

results which are markedly different from the most frequent ones (“fat tails” of the distributions) and which may have a very strong effect on the behaviour of the system.

It is also well-known that power-law distributions of the number of links are frequently observed in biological networks, like e.g. protein-protein networks or gene regulatory networks (Kaneko 2006). In these cases, as well as in many others, the power law concerns the distribution of the number of links per node. The remark concerning the relatively high frequency of far-from-average cases applies also here, and this means that there are some “hub” nodes with a very high number of links, which most strongly influence the behaviour of the network.

Another striking generic property in biology concerns the relationship between the rate of energy consumption (r) and the mass of an organism (m) (West et al. 1997; West 2005). We refer here not to single individuals, but to the average values for a given kind of animal (e.g. cow, mouse, hen, etc.). It has been established by several empirical studies that there is a power-law relationship between the average rate of oxygen intake (i.e. the energy consumption rate) and the average mass: $r = km^{3/4}$ (see Fig. 2.2).

Note that although the mathematical relationship is the same in the two cases above, i.e. a power-law, the semantics is very different. In the first example, the power-law refers to a single variable, and to the frequency of occurrence of a given value in a population, while in the second case it refers to the relationship between two different variables.

What is particularly impressive in the relationship between oxygen consumption rate and mass is that it holds for organisms which are very different from each other (e.g. mammals and birds) and that it spans a very wide range of different masses,

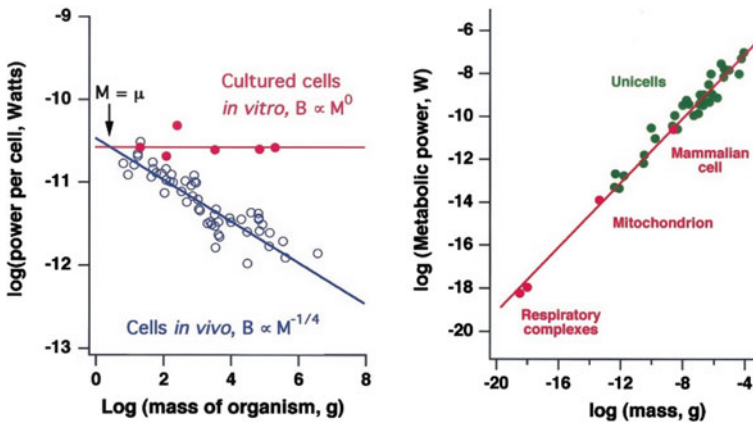


Fig. 2.2 Allometric scaling laws. *Left* the power consumption per cell, when cultured in vitro, is independent of the mass M of the organism it comes from. Since the total power consumption of the organism grows as $M^{3/4}$ (*right panel*) the efficiency of a cell in the organism decreases as $M^{-1/4}$. Comparison with the in vitro behaviour shows that this is a truly systemic property. Reprinted with permission from (West et al. 2002)

from whales to unicellular organisms. Moreover, the same relationship can be extrapolated to even smaller masses, and it can be seen that mitochondria and even the molecular complexes involved lay on the same curve. So the “law” seems to hold for an astonishingly high range of mass values (it has been claimed that no other natural “law” has been ever verified on such a broad spectrum of values).

Of course this is not a law *strictu sensu*, but rather an empirical relationship. It is interesting to observe that an explanation² has been proposed for this regularity, based on the idea that biological evolution has led different organisms to optimize oxygen use and distribution. Indeed, the value of the exponent, estimated from data, is $3/4$, which is surprising, but an elegant proof has been proposed (West et al. 1999) that links the universality of this exponent to the fact that there are three spatial dimensions (and to the hypothesis that evolution works to minimize energy loss).

The two examples discussed above are indeed sufficient to show clearly that generic properties of biological systems, which hold irrespectively of the differences between different organisms, do exist. Let us now consider concepts and models that help us to understand some generic properties.

2.3 Generic Properties of Biological Systems: Concepts

Several candidate (qualitative and quantitative) concepts have been proposed to describe the general properties of complex systems, the second principle of thermodynamics being by far the most successful one. In this section we will briefly mention one of the proposed concepts, that is amenable to at least a partial experimental test, i.e. the notion that evolution should be able to drive biological systems to dynamical “critical” states (Langton 1990; Packard 1988; Kauffman 1993, 1995).

Here “critical” is defined in a specific sense, which is sometimes called “at the edge of chaos” and which somehow differs from e.g. the notion of self-organized criticality (Tang et al. 1988). Dissipative deterministic dynamical systems can often show different long-term behaviours, leading sometimes to ordered states (either constant or oscillating in time), sometimes to quite unpredictable, seemingly erratic wanderings in state space. What is more interesting, is that often the same dynamical system (defined e.g. by a set of differential equations) can behave in one way or another, depending upon the values of some parameters. So there are regions in parameter space where the system is ordered, and regions where it is chaotic. Critical states are those that belong to (or, more loosely, that are close to) the boundaries that separate these regions, so they are close to both ordered and chaotic states.

²This is not the only proposed explanation, but a comprehensive discussion of the origin of allometric scaling laws lies beyond the aim of this book.

It has been suggested (Langton 1990; Packard 1988; Kauffman 1993, 1995; Aldana et al. 2007; Torres Sosa et al. 2012) that critical states provide an optimal tradeoff between the need for robustness (since a biological system must be able to keep homeostasis, notwithstanding external as well as internal perturbations) and the need to be able to adapt to changes. If this is the case, and if evolution is able to change the network parameters, then it should have driven organisms towards critical regions in parameter space.³

This is a very broad and challenging hypothesis, and it can be tested by comparing the results of models of biological systems with data, e.g. models of gene regulatory networks with actual gene expression data. The use of data for this purpose is very different from the more common use of the same data to infer information about the interactions among specific genes. In testing the criticality hypothesis it is instead necessary to look for global properties of gene expression data, like their distributions or some information-theoretic measures (Roli et al. 2011, 2017).

The models to use for comparison should be generic, able to host various dynamical behaviours depending upon the value of some parameter. An outstanding example of this kind is that of the Random Boolean Networks (RBN) model of the dynamics of gene expression. The expression of a given gene depends upon a set of regulatory molecules, which are themselves the product of other genes, or whose presence is indirectly affected by the expression of other genes. So genes influence each other's expression, and this can be described as a network of interacting genes. In RBNs (Kauffman 1969, 1993, 1995) the activation of a gene is assumed to take just one of two possible values, active (1) or inactive (0)—a Boolean approximation whose validity can be judged a posteriori. The model supposes that the state of each node at time $t + 1$ depends upon the values of its input nodes at the preceding time step t . Given that the activations are Boolean, the function which determines the new state of a node is a Boolean function of the inputs.

As it has been anticipated in Sect. 2.1, searching for generic properties requires consideration of ensembles of networks, generated at random (random connections, random Boolean functions) while keeping some parameters fixed (e.g., the average number of connections per node). By comparing experimental data to the properties of ensembles of random networks it is then possible to draw inferences concerning the values of the parameters that define the set. RBNs are indeed dissipative systems that tend to a limited number of different attractors, which represent mutually coherent ways of functioning of the set of genes associated to the nodes of the network; therefore it is straightforward to associate attractors to different cell types.

³Two major variants of this hypothesis have been suggested: (i) that real systems can indeed be in the ordered, more controllable region but close to the critical boundaries, so to be susceptible enough to external changes (Kauffman 1993) and (ii) that in biological systems the notion of criticality has to be taken in a wide sense (Bailly and Longo 2008): while in physical systems one finds critical points, in biological systems one can suppose that they have a finite size. An analogous remark applies as well to critical lines or (hyper)surfaces.

It is then possible to consider the way in which the number of attractors scales with the number of nodes, and to compare it with the relationship of the number of different cell types in different organisms to the number of their genes (Kauffman 1993).

In the so-called *quenched* version of the model, both the topology and the Boolean function associated to each node do not change in time.⁴ The network dynamics is discrete and synchronous, so fixed points and cycles are the only possible asymptotic states in finite networks (a single RBN can have, and usually has, more than one attractor). The model shows two main dynamical regimes, ordered and disordered, depending upon the degree of connectivity and upon the Boolean functions: typically, the average cycle length grows as a power law with the number of nodes N in the ordered region and exponentially in the disordered region (Kauffman 1993). The dynamically disordered region (sometimes called “chaotic”, although of course no real chaos can be observed in finite discrete deterministic systems) also shows sensitive dependence upon the initial conditions, not observed in the ordered case.

One of the most intriguing features of the RBN model is that it allows a distinction between ordered and disordered regimes on the basis of a single parameter, sometimes called the Derrida parameter λ , which depends upon the choice of the Boolean functions and upon the average number of links per node. Ordered states have $\lambda < 1$ and chaotic states $\lambda > 1$; the value $\lambda = 1$ separates order from chaos, and it is therefore the critical value (Kauffman 1993; Serra et al. 2007b).

The technology of molecular biology provides powerful tools to investigate the dynamics of gene expression. In particular, it is possible to analyse the changes induced in the expression levels of all the genes of an organism by knocking-out (i.e., by permanently inhibiting the expression of) a single gene and it is possible to compare the statistical properties of these changes with those of simulated RBNs. The knock-out of a gene can be simulated by choosing it at random among the N nodes of the network and by fixing its value to 0.

It is then possible to compare the time behaviour of the unperturbed (“wild type”, briefly WT) network with that of the perturbed one (“knocked-out”, KO), which is different because of the clamping to 0 of the chosen node (let us call it node R) (Serra et al. 2004b, 2007b, 2015). A node is said to be *affected* if its value in the KO network differs from that of the WT network at least once, after the clamping. Since nodes are connected, the perturbation can in principle spread, and it is not limited to node R, or to those nodes that are directly connected to it. The avalanche associated to that particular knock-out is the set of affected genes, and the size of the avalanche is the cardinality of that set (let us call it ν).

⁴This is of course the most appropriate choice to model a gene regulatory network, where the nodes are the genes and the links represent their mutual influences.

Under the assumptions that the number of incoming links per node A is small ($A \ll N$, where N is the number of genes) and that the overall avalanche is small ($v \ll N$), it can be proven⁵ that the distribution of avalanches depends only upon the distribution p_{out} of outgoing links. In RBNs, the incoming links to a node are drawn at random with uniform probability from the remaining nodes; in this case, the distribution p_{out} is approximately Poissonian and it can be proven that the distribution of avalanches depends only upon the same Derrida parameter that determines the dynamical regime of the network (Serra et al. 2007b). In this case the theoretical distribution is given by Rämö et al. (2006), Di Stefano et al. (2016).

$$p(v) = \frac{v^{v-2}}{(v-1)!} \lambda^{v-1} e^{-\lambda v} \quad (2.1)$$

where $p(v)$ is the normalized probability of finding an avalanche of size v if the Derrida parameter is λ . A comparison with simulations performed on a model RBN with 6300 nodes (the same number of nodes as that of the yeast *S. Cerevisiae*), shown in Fig. 2.3, demonstrates that this expression accurately describes the results of actual simulations of large networks.

It is therefore possible to compare the distribution of avalanches in real organisms to that of model RBNs with different values of the Derrida parameter, and this comparison should tell us whether real cells are critical or not. This is a very interesting example of the way in which simplified models can be used to find generic properties, which cannot be read directly in the data but can be inferred from a comparison between patterns in data and in model results. On the basis of limited data so far available on the yeast *S. Cerevisiae*, it seems plausible to suppose that in that case the network is in an ordered state, not far from the critical boundary (Serra et al. 2004b, 2007b, 2008b; Rämö et al. 2006; Di Stefano et al. 2016). Note that, while this result would rule out truly critical states, it is however one of the possible favourite outcomes of evolution according to Kauffman, i.e. an ordered state close to the critical boundary (see note 3).

However, these conclusions must be taken with some caution: indeed, comparing a Boolean model to continuous data requires the use of some criterion to distinguish affected from non-affected nodes, i.e. to *booleanize* continuous variables. A quantitative criterion can be defined by introducing a threshold θ , so that a node is affected if the ratio of its expression level in the KO network to that of the WT is higher than θ or smaller than $1/\theta$. If the threshold is too small (in the limit $\theta \rightarrow 0$), then one is bound to look just for statistical fluctuations in the expression levels in the two cases, while if the threshold is very high (in the limit $\theta \rightarrow \infty$) no gene appears to be affected. There are heuristic ways to threshold the expression

⁵The assumptions made here are equivalent to supposing that an avalanche never interferes with itself (see Di Stefano et al. 2016 for a precise definition). The non-interference assumption implies that the topology of a spreading avalanche is that of a tree, where each node has a single parent.

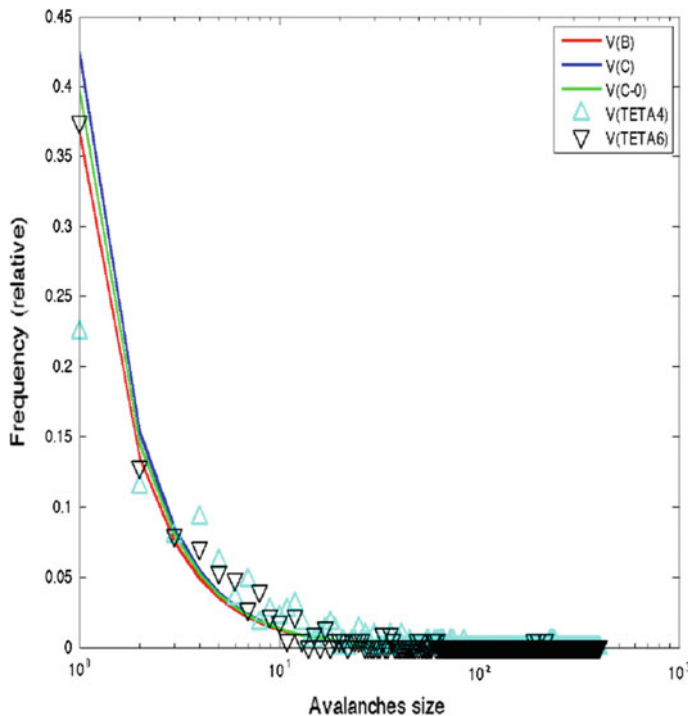


Fig. 2.3 Comparison of the theoretical formula Eq. 3.1 with simulations performed on a network with 6300 nodes, for different values of the Derrida parameter (Di Stefano 2016); the relative frequency is plotted versus avalanche size

values, and the conclusion reported above is based on the use of these heuristic values, which however lack a firm theoretical grounding (Serra et al. 2007b).

Other studies about different biological systems support the hypothesis that the network is either critical or ordered (Shmulevich et al. 2005) or are in favour of the former hypothesis only (Torres-Sosa et al. 2012). Of course further data are needed, but it is nevertheless important to observe that these simplified models can actually open a way to infer very important generic properties of real systems.

The same models also provide relevant evidence in favour of the possibility of successfully applying the RBN model to interpret real biological data. Further model improvements have been developed in order to enlarge the set of possible comparisons with experimental tests (Serra et al. 2004a; Graudenzi et al. 2011a, b) and the effects of cell-cell interaction in tissues (Serra et al. 2008a; Damiani et al. 2008, 2010, 2011; Villani et al. 2006).

Finally, it is worth mentioning that, by taking into account biological noise, the RBN model has been proven able to describe also the main features of cell differentiation (Ribeiro and Kauffman 2007; Serra et al. 2010; Villani et al. 2011,

2013): in this way it has been shown that even such a complex phenomenon can be accounted for by a generic model, without the need of introducing ad hoc genetic circuits.⁶

Let us end this section by stressing again the methodological importance of the approach described here: the model validation is not based upon a direct comparison of the model to the data (like e.g. a direct estimate of the value of a parameter), it rather implies deriving quantitative behaviours from the ensemble of models, and comparing these behaviours to the distribution of values that are actually observed. Finally, this comparison is used to draw inferences about the unknown values of some model parameters.

2.4 What Shall We Model

We will concentrate our modelling efforts on lipid vesicles, which are widely studied as candidate bases for protocell synthesis, although they are by no means the only possibility.

We are aware of the fact that lipid aggregates can have very different morphologies, spanning from unilamellar layers to oligo- or multilamellar membranes (including situations where vesicles contain other vesicles), and can be composed by very heterogeneous materials (Simons and Vaz 2004). Moreover, they can form micelles or vesicles, depending on the chemical environment, on their structure (see Chen and Walde 2010 and further references quoted there) and on packing considerations Israelachvili et al. 1976, 1977).

However in this book we will mainly use the term “vesicle” to refer to a closed structure where a bilayer, formed by amphiphilic molecules, separates an internal water phase from an aqueous external environment (see Figs. 2.4 and 2.5). Indeed, a large part of the experimental efforts thus far have focused on micelles or unilamellar vesicles, made of only one or two components (Chen and Walde 2010). These structures, which are simpler than the multilamellar alternatives, are also more amenable to modelling and will be the target of our models.

One usually speaks of a lipid membrane, although its molecules are indeed amphiphiles, i.e. they display a polar head and a longer lipid tail. The polar heads are found close to the two water phases (i.e. the internal and the external one) while the lipid tails are oriented towards the interior of the membrane. The term liposome is also often used to denote a lipid vesicle.

Different types of molecules are able to form bilayers and also vesicles, including e.g. fatty acids, phospholipids and others. Indeed, some broad reviews exist of the various molecular types that have been proposed (see e.g. Ruiz Mirazo

⁶Of course some hypotheses need to be made; in this case, the key hypothesis is that the level of cellular noise is high in stem cells and decreases during differentiation. There are some experimental indications in favor of this hypothesis, which can and should be subject to further testing.

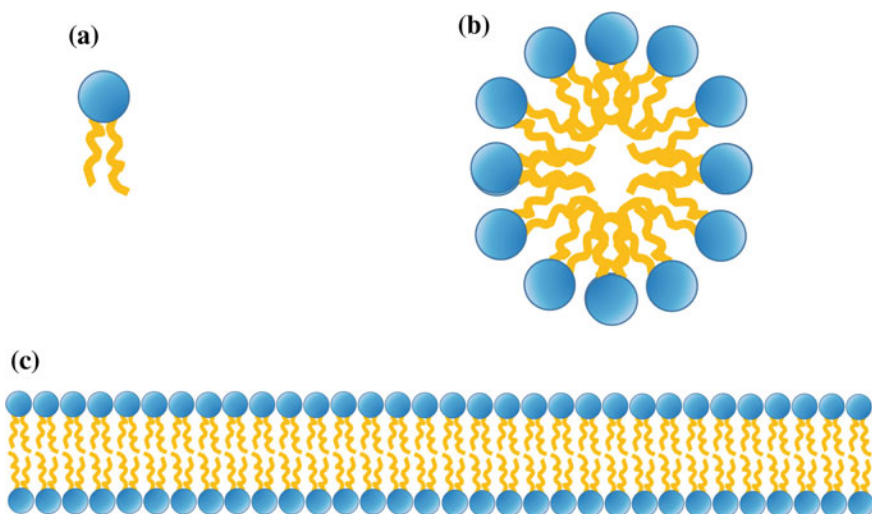


Fig. 2.4 Schematic representation of an amphiphilic molecule (a) and of two energetically favoured supramolecular dispositions, where the lipid tails are separated from the aqueous environment: a micelle (b) and a lipid bilayer, 2D view (c)

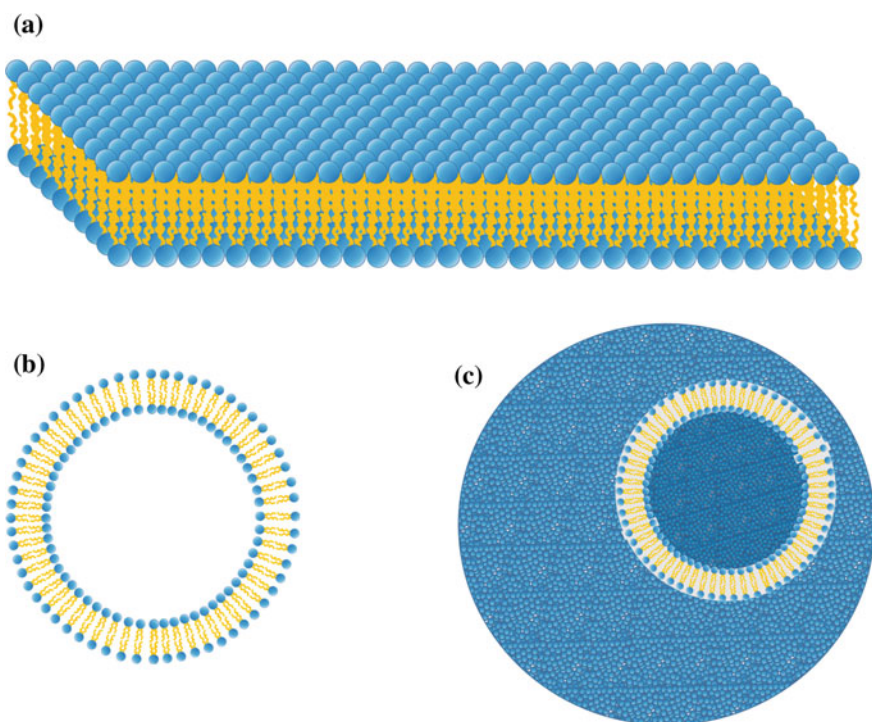


Fig. 2.5 Schematic representation of supramolecular structures. **a** A lipid bilayer, 3D view; **b** a vesicle, 2D view; **c** a vesicle, 3D view

et al. 2014). However, these reviews often contain a lot of detailed information about the actual chemical make-up of the system, which are important for the purpose of reproducing the experimental results, but which are also less relevant for the modelling level we are investigating here.

There are several interesting aspects in vesicles and in protocells that are amenable to dynamical modelling, including:

- the organization of groups of amphiphilic molecules in water or other solvents (McCaskill et al. 2007): they can form different supramolecular structures, like sheets and vesicles, which represent beautiful examples of self-organization phenomena
- the mechanical properties of the membranes (Wang and Du 2008; Alessandrini and Facci 2012)
- the transport processes of different molecular types through the membrane (Wang et al. 2010)
- the intake of amphiphiles in the membrane, their movement and the formation of various domains (rafts) in the membrane itself (Simon and Vaz 2004; Gokel and Negin 2012; Mc Connell and Vrljic 2003)
- the description of the process of fission, where a single vesicle splits into two vesicles undergoing changes in shape (Luisi et al. 2004); this is a complex phenomenon, which requires a change of shape and the subsequent breaking of the channel connecting the two parts of the parent vesicles. Beautiful studies include those of Svetina (2009), Morris et al. (2010)

A major issue concerns the most appropriate modelling level. Protocells are made out of molecules, so models dealing with molecular properties can be important, ranging from the level of quantum chemistry to that of molecular dynamics. However, the description of the properties of vesicles and protocells typically require a coarser graining than those of the previous approaches, which are well suited to deal with single molecules (perhaps in a heat bath) or with few interacting molecules. An interesting set of models based upon the DPD (Dissipative Particle Dynamics) approximation has also been studied (Fellerman et al. 2007).

Dealing with supramolecular structures like protocells, we will mostly ignore the details of the molecular level. In our models the behaviour of the amphiphiles and of the proto-genetic molecular species will be described by the methods of chemical kinetics. Both deterministic and stochastic models will be considered: the former are more amenable to theoretical treatment and to fast simulations, while the latter are required to deal with cases where there are only few copies of some important molecular types. The models of Chap. 3 are essentially deterministic, while intrinsically stochastic models will be studied in Chaps. 4 and 5.⁷

⁷Note however that in complex systems science it is sometimes convenient to consider different models of the same phenomenon; so, in Chap. 3 the effects of random fluctuations will also be explored, and in Chaps. 4 and 5 some deterministic approximations will also be used whenever appropriate.

There are also other interesting topics in protocell research, but in this volume we will focus our attention on the coupled processes of replication of the “genetic” molecules and of the growth and duplication of the lipid “container”, and we will try to uncover some generic features of these processes, as discussed in the previous sections of this chapter.

In the following chapters we will therefore take a very simplified view of a protocell: the majority of the models that will be described and analysed are fairly abstract, and do not make explicit reference to the specific properties of the membrane. What is required is that (i) closed compartments form spontaneously (ii) their membranes are selectively permeable to some but not to all the chemicals and (iii) they are able to grow and to fission when a certain critical size has been reached. While lipid vesicles are the best known systems of this kind, other vesicle-forming chemicals could do the job, including micelles (Mitchell and Ninham 1981), reverse micelles (Pileni 1993), lipid droplets (Thiam et al. 2013) and others.

Note also that protocells might be used in several interesting applications, including intelligent drug delivery, recognition of other protocells or of some other “agent” in the body or in the environment, information processing and many more else. These applications might be one of the main reasons of interest for protocells, but their treatment also lies beyond the scope of this work.

Modelling Protocells

The Emergent Synchronization of Reproduction and
Molecular Replication

Serra, R.; Villani, M.

2017, XV, 182 p. 46 illus., 33 illus. in color., Hardcover

ISBN: 978-94-024-1158-4