Modeling collective behaviour in biology: Computational approaches

A thesis submitted to the University of Hyderabad,
in partial fulfillment of the requirements for the award of the degree of
Doctor of Philosophy

Donepudi Raviteja
Reg. No: 12LTPH11
Department of Biotechnology & Bioinformatics
School of Life Sciences, University of Hyderabad
Hyderabad - 500046
June 2018

Dedicated to

Parents
and
University of Hyderabad

Declaration

I hereby declare that the work carried out in this thesis entitled "**Modeling collective behaviour in biology: Computational approaches**" is entirely original. It was carried out by me in the Department of Biotechnology & Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad. I further declare that it has not formed the basis for the award of any degree, diploma, membership or similar title of any university or institution.

Donepudi Raviteja
Department of Biotechnology & Bioinformatics
School of Life Sciences, University of Hyderabad
Hyderabad, Telangana, India

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisors Prof. Ramakrishna Ramaswamy and Prof. Niyaz Ahmed. This thesis would not have been possible without their continuous and uninhibited support.

Prof. Ramaswamy, with his immense knowledge and composed character, has always been a great inspiration. His never say no to a discussion with student attitude and the enthusiasm to participate in diverse scientific topics, allowed me to explore a multitude of scientific concepts. His participation at all stages of research, and his optimism was of great help and gave me confidence. The understanding nature and unwavering support of Prof. Niyaz Ahmed, had made my PhD journey possible. His insistence of including me into meetings at PBL allowed me to have a second lab at Life Sciences. I find myself extremely favoured to have them as my supervisors.

I would like to thank my doctoral research committee members, Profs. Sharmista Banarjee and Debashish Barik, for their comments and suggestions. I would like to thank, Prof. Rob de Boer and Dr. Aridhaman Pandit for hosting me at the University of Utrecth. I would like to thank Prof. Anantha Krishnan for facilitating my first meeting with Prof. Ramaswamy and motivating me to pursue the PhD.

I am grateful to my mother, father and sister Aruna, Sambasiva Rao and Bindu, who have provided me all possible support in my life. I am also grateful to my other family members and friends who have supported me along the way.

I thank CSIR for providing me with Junior and Senior research fellowships (SPMF) for the duration of my PhD.

I am thankful to all my fellow students and colleagues at the NonLinear Dynamics lab (JNU) and Pathogen Biology Laboratory, with a special mention to Vidya, Sangeeta, Nirmal, Rupesh, Shakir, Suraj, Pankaj, Anagha, Haider, Amit, Emil, Umesh Sir, Samir,

Sharukh, Ramani, Sumeet, Aditya, Shankar, Arif, Narender, Sabiha, Arya, Priyadarshini, Kishore. It was fantastic to have the opportunity to meet with you and have interesting and long discussions.

The faculty at School of Life Sciences (UoH) have always been a source of inspiration and I thank them for their teaching and academic support.

I am also grateful to the all the administrative and non-teaching staff at University of Hyderabad and particularly Rajendra Gowd, Krishna Ram, Raju Bhayya, Raju, Venkat, Arun, and Rahul for their unfailing support and administrative assistance.

Thanks for all your encouragement!

Publications

1. *The collective dynamics of NFκB in cellular ensembles: Cluster synchrony, Splay states, and Chimeras*
   R. Donepudi and R. Ramaswamy.
   EPJ ST Special Issue: Nonlinear effects in life sciences, (2018). (Accepted).

2. *By-product Group Benefits of Non-kin Resources Sharing in Vampire Bats*
   R. Donepudi and R. Ramaswamy.
   IOP Journal of Physics: Conference Series, ICCSE 2017 (2018). (Accepted).

3. *Modeling Long Lifespans in Eusocial Insect Populations*
   R. Donepudi and R. Ramaswamy. (Under review, 2018).

4. *Eukaryotic genome expansion: A consequence of asymmetric opportunity costs of genome expansion between proto-mitochondrion and the proto-nucleus?*
   R. Donepudi and R. Ramaswamy.
   To be communicated.

Contents

CHAPTER 1

Introduction

In a discursive article that appeared in the journal *Science* in 1945, A.B. Novikoff [1] elaborated on the applicability of the concept of integrative levels to biology. The fact that biological systems have a hierarchy of organisation levels that can be differentiated by size, function, or complexity is apparent. At the sub-cellular level, the molecules of biology are important, at the cellular level, the different structures such as organelles or the nuclei, and so on to levels such as the tissue, that looks at collections of cells, to organs, and further to the entire organism, an organization of organs, and on to populations that comprise of organisms, to communities, namely groups of populations, to the ecosystem and eventually to the biosphere itself [2]. This organizational scheme incorporates — but is also beyond — the Linnaean classification of all living organisms: since the number of levels is large, this hierarchy is effectively continuous, but the transitions from one level to the other is abrupt, and thus the hierarchy is also discontinuous [1, 3].

The idea that all matter can be thought of in terms of integrative levels can be useful. At each stage, the individual units of lower levels are coarse-grained, in the sense that they are organised and integrated into a single system over which the complexity of newer levels are superimposed [1]. As the construction of newer levels is a matter of choice and can therefore be an unending process, one level in its entirety will always be part of another [4]. Within such a classification, though each level is discrete and possesses unique properties of structure and behaviour.

These notions have great relevance in contemporary approaches to systems and integrative biology [5, 6]. In this thesis, I have addressed a set of biological problems at various levels of the overall organization, ranging from the subcellular to the cellular to

the organismal and the social. The studies presented use various multi-level modelling schemes. Mathematical and computational models are handy tools in the analysis of a specific scientific problem where it is crucial to isolate parts of the whole, as it is to integrate these very parts into the structure of the whole [7, 8, 9, 10]. Successful application can result in explanations of a range of natural phenomena, with verifiable quantitative and qualitative predictions.

In this Chapter, I describe the different modelling techniques that have been used in the course of my studies, and briefly introduce the problems that I have addressed. The underlying theme is that of multilevel modelling and collective behaviour, with features of each level, although due to their constituent parts, appearing only when they are formed as a new system. Therefore for a complete understanding of a given level, knowledge of the laws of the lower levels are necessary but might not be sufficient. The laws defining unique properties of each level are qualitatively distinct and to understand them appropriate research methods at that level are required [1, 3, 4]. The organisation principles, which include the reciprocal relation of parts with each other and of the parts to the entire system, are defined by these laws [4].

1.1 Mathematical modelling

The mathematical modelling of biological problems has a long history [11]. A major aim of the quantitative approach to any discipline is the ability to make reasonably accurate predictions based on empirical observations, and in this process, a mathematical model is essential. At the same time, since a mathematical model is an attempt to reconstruct the mechanisms of processes that may occur at vastly different length- and time-scales, it should not be forgotten that models are, *per se* not explanations, and cannot provide complete solutions to most biological problems.

An additional feature of many models is that they are perpetually in a state of being "under construction", in the sense that the process of improving them by the incorporation of more sophisticated observations and the use of better computational and mathematical tools is an ongoing one. To start with, most models are minimal, in the sense that Ockham's razor [12] is a very useful guiding principle, and only gradually acquire the complexity that becomes necessary as more and more data becomes available.

A number of different biological problems have been usefully modelled at both qualitative and quantitative levels. These range from the classic work of Lotka and Volterra [13, 14] that introduced coupled differential equations to model a predator-prey system in population ecology, to the Turing model of reaction-diffusion systems [15] for biological pattern formation or the classic work of Hogkin and Huxley [16] on cell membrane action

potentials. The examples are many, as are the different tools that are used, ranging from discrete difference models [17] to differential equations in several dimensions, cellular automata [18], game-theoretic models [19], genetic algorithms [20], and agent based modelling [21].

An early instance of a highly successful application of mathematics to biology was to the study of population growth. Consider the following equation for the growth rate of bacteria in a medium with infinite nutritional resources. If $p(t)$ denotes the population at time t, and the rate of fission of bacteria is a_k per minute, then the rate of change of the population is given by Malthusian growth [22], namely

$$\frac{dp}{dt} = a_k p. \tag{1.1}$$

This exponential growth function (see Fig. 1.1) is an example of linear equation, and the solution is straightforward, giving exponential growth, namely $p(t) = p(0) \exp a_k t$, where $p(0)$ is the value of the population at time $t = 0$.

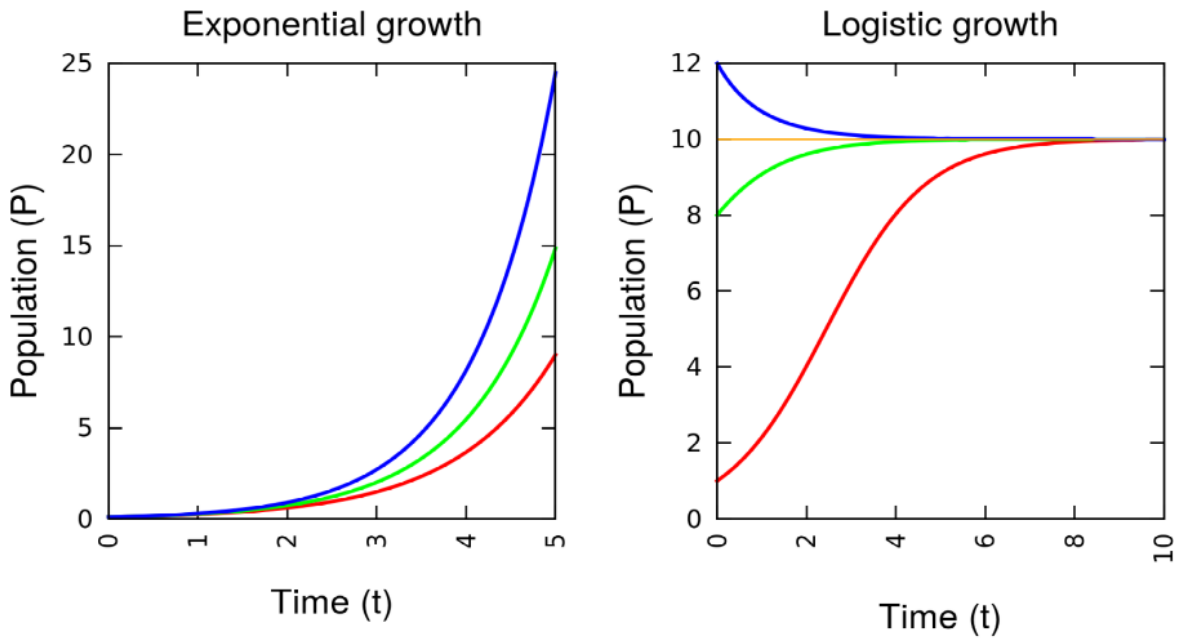

Figure 1.1: The population growth for three different values of the growth rate, $a_1 < a_2 < a_3$ shown by the red, green and blue curves respectively in the case of exponential growth (left panel). For logistic growth (right panel) however, the population saturates to the value K regardless of initial condition (three instances are shown).

A modification of this growth law was introduced by Verhulst in 1838 [23]: this is the logistic growth law, a nonlinear equation where the population has a limit, the "carrying

capacity" K.

$$\frac{dp}{dt} = a_k p(1 - \frac{p}{K}) \tag{1.2}$$

In the above differential equation, the population is treated as a continuous variable. When the population size is very large, the variation may be taken as essentially continuous. For small populations of organisms though $p(t)$ can only take integer values, with changes occurring at definite times, and different modelling strategies need to be used. Solutions of this equation can also be obtained in a straightforward manner,

$$p(t) = \frac{Kp(0)e^{a_k t}}{K + p(0)(e^{a_k t} - 1)}$$

and as $t \to \infty$, it is clear that $p(t) \to K$; see Fig. 1.1.

1.2 Dynamical systems

In order to model a physical system that comprises a set of units or components (which can be particles, molecules, or other entities), one needs a set of *state* variables that are needed to completely specify the state of the system. The evolution of the state can often be specified by a rule, namely a function that governs the change [24, 25]; together these constitute the dynamical system. If there are n state variables which are denoted by the (n-dimensional) vector $\mathbf{x}$, and $\mathbf{F}$ denotes the evolution functions for the different state variables, then one has

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}, \mu, t) \tag{1.3}$$

where μ denotes the parameters of the problem. Note that higher order differential equations can be replaced by a larger number of first-order differential equations through a fairly straightforward procedure. Furthermore, if the dynamical system is autonomous then the functions $\mathbf{F}$ do not depend explicitly on time. Non-autonomous systems can be converted, by introducing an additional variable, to an autonomous one, so it suffices to consider coupled first order ordinary differential equations. For dynamical systems with discrete time the dynamics are governed by coupled difference equations or iterative maps.

An important characteristic of a given dynamical system is in the evolution of phase space volumes under the dynamics. If the volume remains unchanged as the system evolves it is said to be conservative. On the other hand if the evolution equations are such that volumes shrink as a function of time, then the system is said to be dissipative. A distinction, therefore, between the two kinds of systems, is that one can have different asymptotic dynamics; in particular, there can be attractors in dissipative systems, as

discussed below.

In dynamical systems, the sensitivity to initial conditions or the rate of separation of arbitrarily close trajectories is quantified by the Lyapunov exponents [26, 27, 28]. The number of Lyapunov exponents is equal to the dimension of the phase space, and if the largest (or maximal) Lyapunov exponent (MLE) is positive, the dynamics shows sensitivity to initial conditions, namely chaotic motion. If the distance between two nearby trajectories at time t =0 is δx_0), then this distance grows at a rate

$$|\delta x(t)| \approx e^{\lambda t}|\delta x_0|, \tag{1.4}$$

where λ is the MLE. An estimate of the largest Lyapunov exponent is given as

$$\lambda = \lim_{t\to\infty}\lim_{x_0\to 0}\frac{1}{t}\log\frac{\delta x(t)}{\delta x_0}. \tag{1.5}$$

Equivalently, one can compute the Lyapunov exponents in terms of the eigenvalues of the Jacobian matrix, evaluated along the trajectory [25] or by other numerical procedures [29].

1.2.1 Attractors, Chaos, and Bifurcations

In a dissipative system, since volumes contract, any set of initial conditions will, with the passage of time, contract into a subset of zero volume, namely a lower dimensional object. It may also happen that this final state is the same, regardless of initial conditions: such a lower dimensional object is termed an *attractor* of the dynamics. The attractor is an invariant set, in the sense that any trajectory which starts on the attractor will stay on the attractor.

Simple examples of such attractors are fixed points or limit cycles (see Fig. 1.2): after a transient period, the trajectory reaches the attractor and is then unchanged (on the fixed point) or shows periodic motion (on the limit cycle).

Other geometries are possible. The attractor may be a torus, and on this the motion will in general be quasiperiodic rather than periodic. Of most interest, though, is when the attractor has a fractal geometry, and in such cases, the motion cannot be periodic or quasiperiodic: such attractors are termed *strange* (see Fig.1.2). A chaotic strange attractor is a strange attractor on which nearby points diverge from each other while remaining on the attractor [26, 30], while a strange nonchaotic attractor does not have sensitivity to initial conditions [31].

When the parameters of a dynamical system are varied, the motion can undergo qualitative changes that are termed bifurcations [30, 32]. These changes occur in distinct

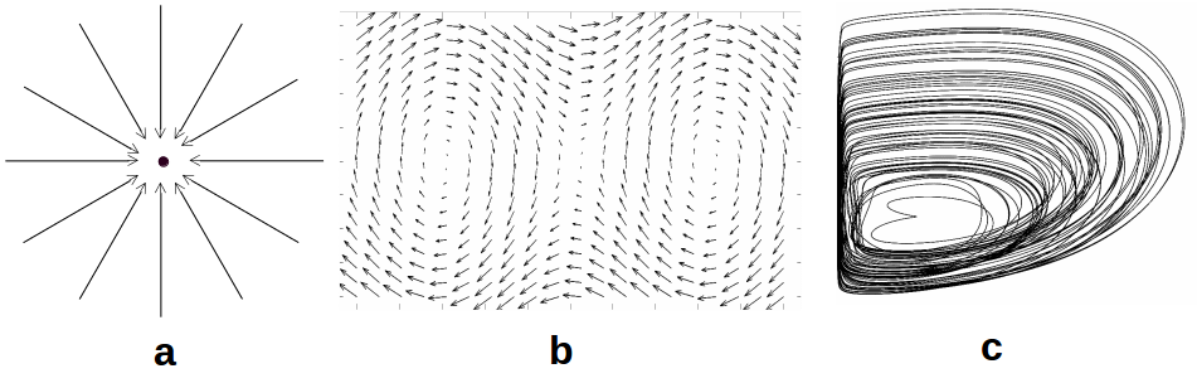

Figure 1.2: Phase portrait of 2 dimensional attractors: a) Fixed point for logistic growth, b) Limit cycles in a pendulum, and c) a chaotic strange attractor in $Nf\kappa B$ oscillations (see Chapter 5).

and characteristic ways, and typically affect the stabilities of invariant sets such as fixed points or periodic orbits. Some of the most common bifurcations can be seen in the iterative map $x_{n+1} \rightarrow rx_n(1 - x_n)$ when the parameter r is varied (see Fig. 1.3). At $r = 1$, for instance, a stable fixed point collides with an unstable fixed point and their stabilities are exchanged: this is termed a *transcritical* bifurcation. At $r = 3$, there is a *period–doubling* bifurcation: a stable fixed point becomes unstable and simultaneously, a stable orbit of period 2 is created. In general, an orbit of given period becomes unstable and an orbit of twice that period is created at all period doubling bifurcations. At a *saddle-node* bifurcation point, a pair of fixed points, one stable and one unstable are simultaneously created, and at $r = 1 + \sqrt{8}$ a period-3 orbit is born via such a bifurcation in the logistic map. Repeated period doubling bifurcations lead to the so-called period-doubling cascade, finally resulting in chaotic dynamics (see Fig.1.3).

For local bifurcations, the qualitative changes are confined to local stability of attractors while for global bifurcations, large invariant sets collide with other attractors and cause changes in the topology of trajectories such that they are extended to arbitrary large distances. Some of these are shown schematically in Fig.1.4. In addition to the ones described above, there can be pitchfork bifurcations characterized by one fixed point becoming three as shown in c). At a Hopf bifurcation shown in d), a periodic orbit is created from a fixed point which becomes unstable [33].

It is also possible that for a given value of the parameters, there can be more than one attractor that is stable: this is termed *multistability* [34]. When this happens, it is important to interrogate the dependence of the final state on the initial condition. This defines the basin of a given attractor: this is the set of initial conditions that eventually lead to that attractor. When there are several attractors, the basins can be arranged in

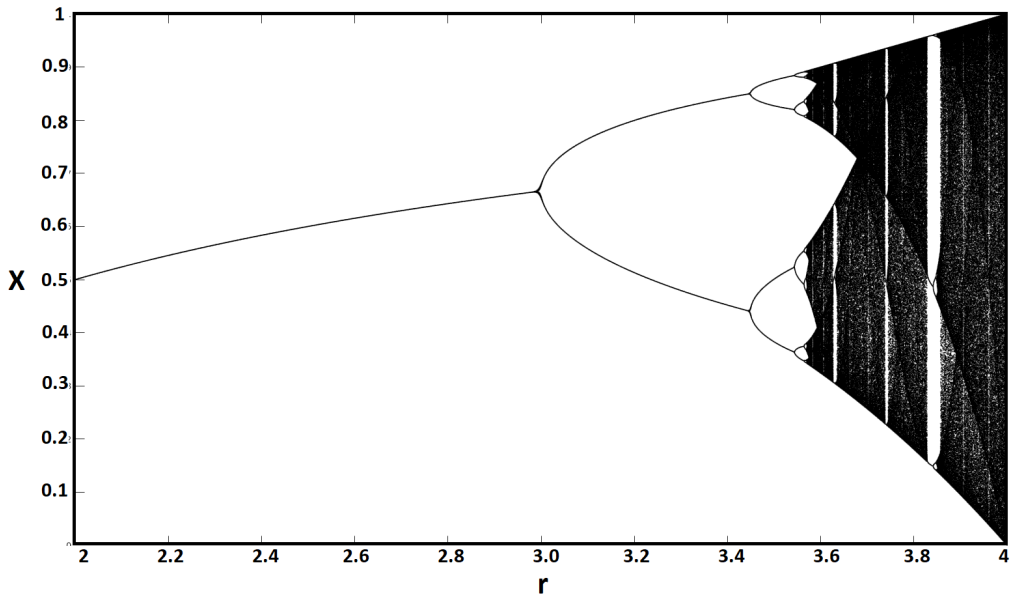

Figure 1.3: Bifurcation diagram for the logistic map as a function of r. See the text for a discussion of the various prominent features.

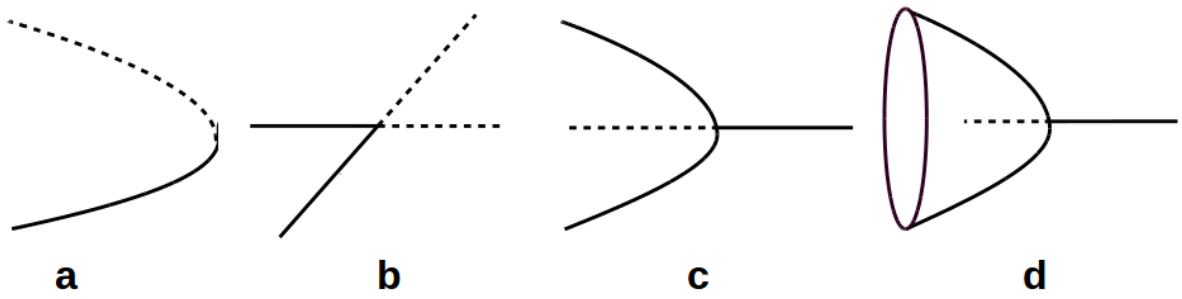

Figure 1.4: Schematics of a few bifurcations discussed in the text. a) Saddle-node bifurcation, b) Transcritical bifurcation, c) Pitchfork bifurcation, and d) Hopf bifurcation. Stable orbits are depicted through solid lines, and unstable ones through dashed lines

a complicated way in the phase space. The boundaries are often not smooth, especially when there is more than one chaotic attractor [35], and furthermore, the basins can be riddled or intermingled [36], making the system sensitive not only to initial conditions, but to the final state as well [35]. In biological systems with noise the system can hop between various attractors, producing a much richer behaviour [37].

1.2.2 Coupled Oscillators

Natural systems are rarely isolated: it is necessary to consider connections between them and to understand the effects that arise due to couplings between them. For biological systems, for instance, depending on the level [1] at which the analysis is being done, this is very often the case: the individual subunits are usually connected with one another, or are coupled indirectly through the environment. These interactions make it necessary to consider the overall system as consisting of an ensemble of subunits, and to investigate the novel emergent behaviour that arises due to the coupling.

Coupling interactions can be of various types. The skew–product or *master-slave* coupling is unidirectional, with a master system driving an enslaved unit, without being affected by it. When the coupling is bidirectional, the two systems are affected by each other. Depending on the way in which each system is coupled with all others within an ensemble, there are three types of coupling: (i) Global coupling occurs when all systems are connected to all other systems as in Fig. 1.5 (a), (ii) Local coupling, when systems are connected to only via their nearest neighbours; see Fig. 1.5 (b), and (iii) Nonlocal coupling, when the strength of the coupling depends on the distance between the systems; see Fig. 1.5 (c).

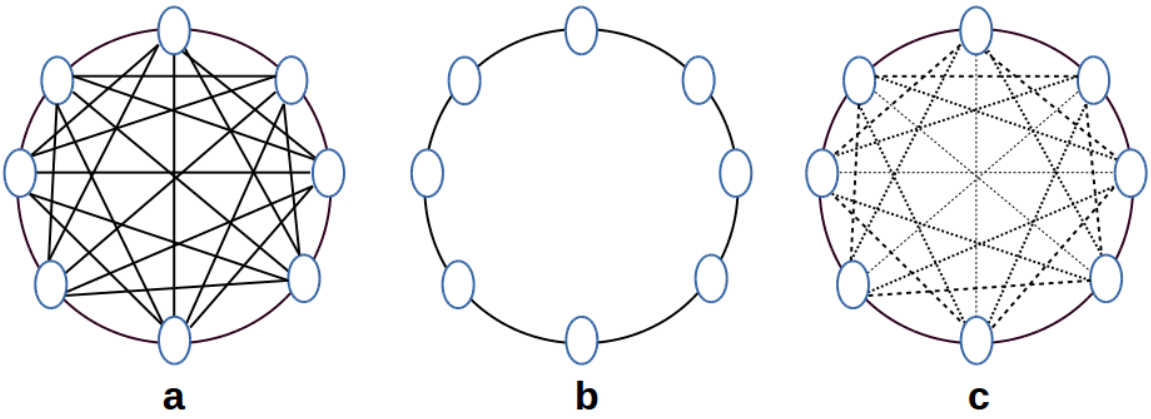

Figure 1.5: Types of bidirectional coupling, a) global, b) local, and c) non-local coupling. In (c), the coupling strength is dependent on the distance, here indicated by different types of lines.

When an ensemble of oscillatory dynamical systems are coupled locally or globally, there can be interesting emergent behaviour. The most striking is the phenomenon of global synchronisation which is depicted in Fig. 1.6 (a) namely the adjustment of rhythms of oscillating objects due to the interaction. There are, by now, a very wide range of behaviours that are classified as various types of synchronization, ranging from complete synchrony, when all the variables of the subsystem are identical to one another, to generalized synchrony, when there is only a functional relationship among the variables of the different units [38]. For purposes of this thesis, I would like to draw attention to the possibility of splay states (see Fig. 1.6 (b)) when the variables of each of the oscillators has a phase lag with respect to the others. The phenomenon of cluster synchronisation (see Fig. 1.6 (c)) is when distinct subgroups of oscillators are synchronised to different limit cycles. Dynamical chimeras (see Fig. 1.6 (d)) occur when both coherent and incoherent behaviours arise in a group of identically coupled identical oscillators. This is a form of symmetry breaking that is beginning to be understood in some detail [39].

1.3 Evolutionary dynamics of population models

I now switch gears to describe another modelling protocol that I have used in this thesis, namely evolutionary game theory [40, 41] which is a mathematical and computational approach to understanding the evolution of biological populations [19, 42, 43, 44]. The individuals that constitute the population are termed players, and the interactions between them is defined in terms of a formal game. Behavioural phenotypes on the basis of which the players take action in the game are termed strategies. When competing via natural selection, different types of players reproduce to the extent of their ability: this defines *fitness*. Reproduction can either result in an exact copy, or in advanced models, be a variation due to mutations. In general, the average fitness of a type of player depends on its own strategy, that of other types, and on their frequencies in the population [44].

In a typical evolutionary dynamics simulation, the players (which are biologically stipulated) are randomly drawn from a large population and the game is played repeatedly between them. Each player multiplies depending on its advantage or "pay-off". The evolutionary selection process will operate and leads to dynamics in the distribution of various types of players (strategies). The dynamics can lead to a stable state with one of the strategy wiping out others, it can lead to equilibrium state where the frequencies of each type are fixed (coexistence), or can also lead to non-equilibrium dynamics like stable cyclic behaviour or chaos [19].

In order to allow a new strategy to emerge in a given population, the traditional game

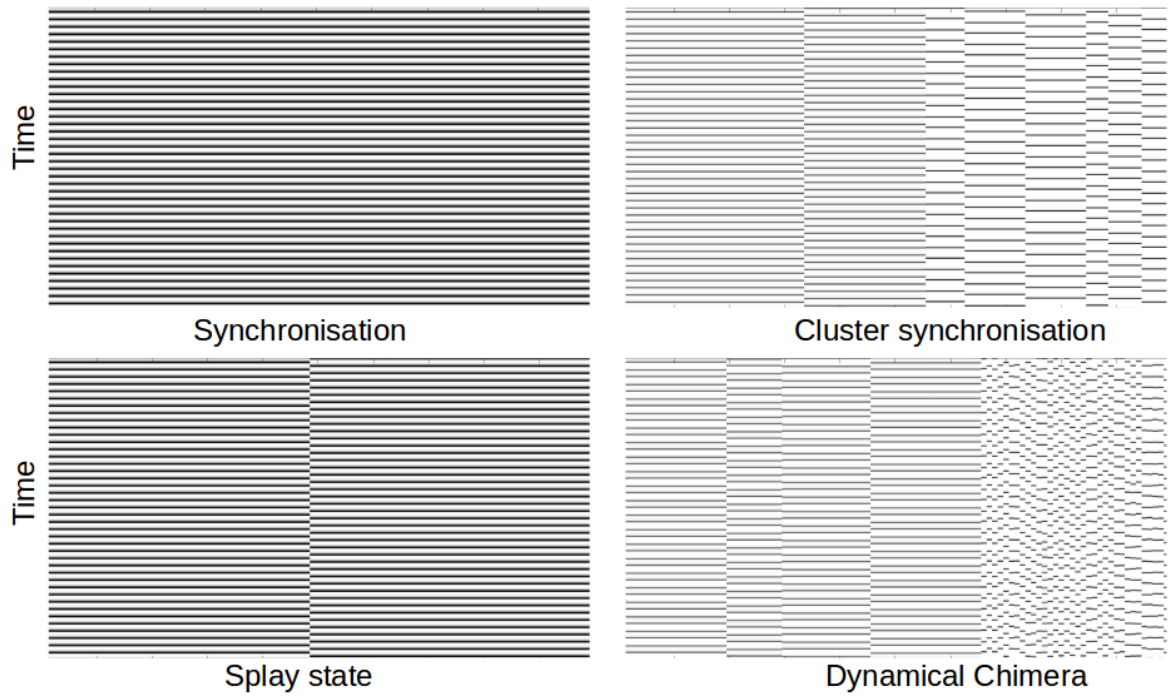

Figure 1.6: Different collective behaviour that can arise in a population of coupled nonlinear oscillators. Shown above are spacetime plots for a set of N=64 coupled oscillators, where one of the variables is chosen for display. The different oscillators are arranged along the x-axis, while time plotted on the y. The chosen variable is converted to grayscale so that the maximum is darkest and the minimum is white. In the upper left panel all oscillators behave identically, in a state of complete synchrony. In the lower left panel, the oscillators separate into two sets which are exactly out of phase with one another, in what is termed a splay state. The upper right panel shows cluster synchronization, namely the system separates out into a number of groups of oscillators that bear no specific phase or amplitude relationship with each other, but all the oscillators within a cluster are in perfect synchrony. In a chimera state, as shown in the lower right panel, there are clusters where the synchrony is destroyed and one has an asynchronous cluster.

theory needs to be extended since the usual theory [19] is dominated by static models wherein the attempt in a so–called rational game is to find a best or unique solution, the Nash equilibrium that comes from an evolutionary stable strategy [41, 43]. This is the state where each player has complete information about all other players and rationally chooses the best-strategy, such that no player can improve their payoff by switching only their own strategy [45]. The concept of an "unbeatable" stable strategy assumes an underlying population dynamics which in turn depend on the structure of population, mechanisms of inheritance (namely the transmission of traits), and the time scales of the various processes. Evolutionary models with dynamics explicitly incorporate such dependencies [19].

1.3.1 Growth models

One can incorporate game-theoretic strategies in the deterministic growth models discussed earlier to give an example of evolutionary strategies. Consider exponential growth of an asexual population, given by Eq. (1.5) where there are two strategies given by two different values of the growth rate a. Denote these by a_1 and a_2, with $a_1 > a_2$. If p_1 and p_2 are the populations of the agents adopting these strategies, respectively, one can easily see that the ratio between "size" of the two strategies, namely $R \equiv p_1/p_2$ has the rate of change

$$\dot{R} = \frac{\dot{p_1}p_2 - p_1\dot{p_2}}{p_2^2} = (a_1 - a_2)R, \tag{1.6}$$

leading therefore to the selection of the type with larger growth rate, namely a_1 which will be the only surviving type (see red curve in Fig. 1.7). For the logistic model, namely

$$\dot{x} = r_1 x(1 - \frac{x+y}{K})$$

and

$$\dot{y} = r_2 y(1 - \frac{x+y}{K}),$$

where the dominant strategy will saturate at the carrying capacity K. Depending on the initial conditions, coexistence of both types can be expected (see the blue curve in Fig. 1.7).

This can become a simple evolutionary dynamical model if one incorporates natural selection. Assume that the players reproduce at different rates,

$$x_{t+1} = r_1 x_t^a (1 - v) \tag{1.7}$$

$$y_{t+1} = r_2 y_t^b (1 - v) \tag{1.8}$$

$$v = x + y \tag{1.9}$$

Given $a = b \geq 1$, the dominant strategy is the one with greater r_i. The dynamics of the dominating strategy is equivalent to a single strategy population (compare Figs. 1.8 and 1.3) and can show fixed points, limit cycles and chaotic dynamics. For $a, b < 1$ there can be coexistence and depending on parameter values both populations can either show fixed point, limit cycle, or chaos (see right panel of Fig. 1.8).

1.3.2 Quasispecies and the Replicator equation

A standard tool that is used to analyse biological games where the fitness of strategies is frequency dependent is the replicator equation [19, 44]. The replicator equation allows for the fitness function to include frequencies of different types of strategies, in contrast

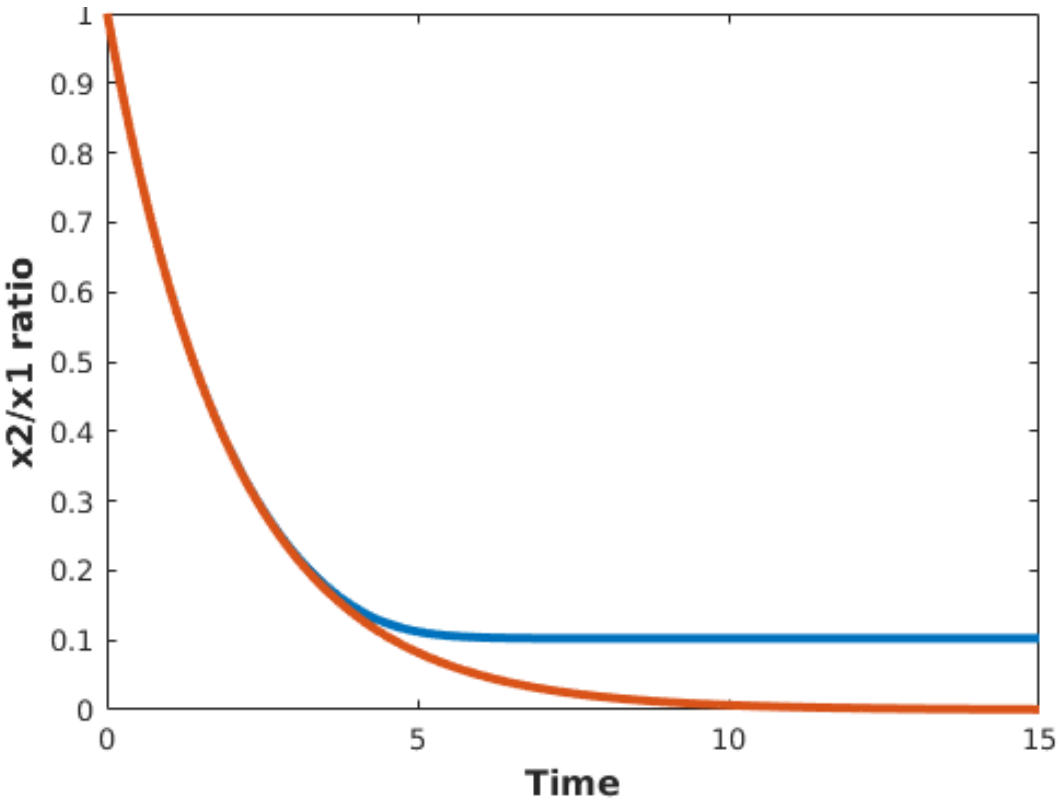

Figure 1.7: Population ratio as a function of time for competing populations. For exponential growth (the red curve) the population ratio goes to zero, namely $p_2 \to 0$, signifying extinction of species 2. For logistic growth with $r_1 = 2$, $r_2 = 1.5$ and $K = 10$, the blue curve asymptotes to a nonzero value as discussed in the text.

to the constant values taken in the previous examples. Fitness functions can be linear if the payoff is from pairwise encounters, or nonlinear if large populations are considered. Several strategies can coexist in the steady state, and beyond this, fluctuating frequencies and chaotic dynamics are also possible [19]. The general form of a replicator equation is given by

$$\dot{x_i} = x_i[f_i(\mathbf{x}) - \phi(\mathbf{x})], \tag{1.10}$$

$$\text{where } \phi(\mathbf{x}) = \sum_{j=1}^{n} x_j f_j(\mathbf{x}). \tag{1.11}$$

The fitness f_i of a player of type i is a function of frequencies of all the strategies ($\mathbf{x} = [x_1, x_2, \ldots, x_n]$), and $\phi(x)$ is the average fitness of the population which is here normalised to 1.

In the above replicator equation, I assume that the strategies are competing genotypes and during reproduction the genotypes produces accurate copies. The genotype with

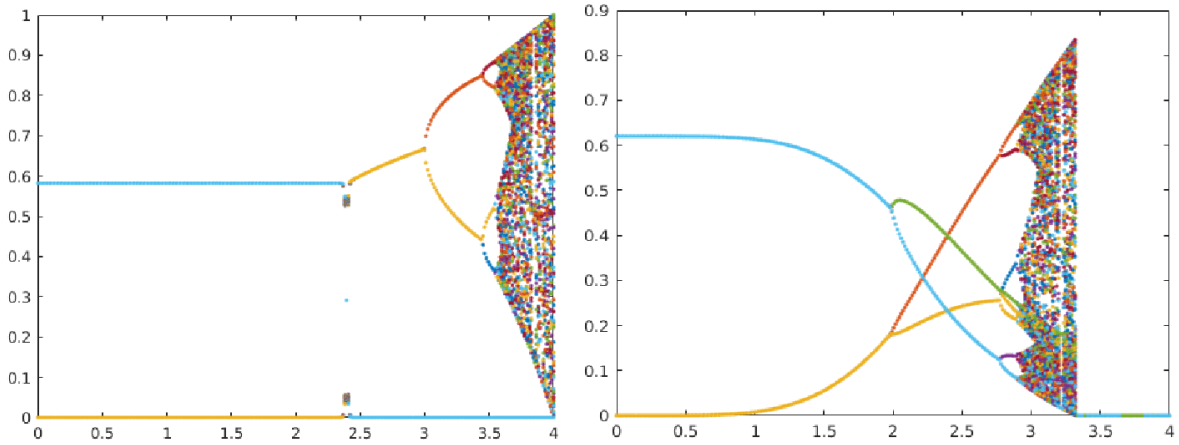

Figure 1.8: Superimposed bifurcations of competing logistic population models (maps). For the left panel $a = b = 1$, $r_1 = 2.4$. The abscissa corresponds to r_2 in both panels. On the right, I take $a = b = 0.8$, $r_1 = 2.4$. When both populations grow subcritically, namely $a, b < 1$, there is coexistence, eventually leading to extinction.

greater fitness will out-compete others and establish itself as the dominant strategy. It can also happen that two genotypes can coexist in steady state or oscillate. Hence the selection can be seen as acting on individual genotypes (species) each of which has a specific fitness.

A quasispecies is a distribution of related genotypes (mutations) that exists in a high mutation environment and are generated by a mutation-selection process [44, 46, 47]. In a mutation-selection process the genotypes are mutated while replicating and produces genotypes which replicate at different rates. The genotypes that replicate at a faster rate will increase in the population, while those that replicate more slowly will reduce in number. Since slower genotypes can be produced by mutation of faster genotypes, they can be constantly replenished. When the relative loss of slower genotypes is equal to their production by mutation, an equilibrium is established. Due to high mutation rate it is more relevant to discuss the fitness of a the entire population, namely the quasispecies as a whole

The mutation-selection process of a quasispecies on a constant fitness landscape is given by following deterministic equation,

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j q_{ji} - \phi(\mathbf{x}) x_i \tag{1.12}$$

where q_{ji} is the probability of a mutation converting genotype j to genotype i. The other quantities are as in the replicator equation. In simulations of populations when both fre-

quency dependent selection and mutation are important, the replicator and quasispecies equations can be merged into a single equation,

$$\dot{x_i} = \sum_{j=1}^{n} x_j f_j(x) q_{ji} - \phi(\mathbf{x}) x_i. \tag{1.13}$$

Here as for replicator equation, the fitness f_i of a player of type i is a function of frequencies of all the strategies ($\mathbf{x} = [x_1, x_2, \ldots, x_n]$.)

1.3.3 Evolving populations

In studies of evolutionary population dynamics, there is the assumption that after sufficient time there is a a well mixed population spatially as well as temporally, and with the interaction between all players being equiprobable. This is not always valid, given that in most simulations there are inhomogeneities. This poses a limitation in studying realistic many-player models wherein the interactions between players depend on space, time and population structure.

It is possible to modify the evolutions equations to include the structure of the population, to include reaction-diffusion terms, and also to incorporate various types of lattice geometry, neighbourhood structure and updating rules. Recent models of evolutionary dynamics incorporate graph theory [48] to model asymmetric interactions. Age structure in populations can also be included: this would make it possible to have time-dependent growth rates or death rates as is commonly observed in natural populations [11],

Evolutionary dynamics models can thus be extended to incorporate complexities [49] and in the work presented later in this thesis, I have explored one or more of these extensions.

1.4 Agent Based Modelling

Agent based (or individual based) modelling (ABM) is a simulation strategy that employs autonomous agents, objects, or entities. Agents are characterised by a rule-based decision-making and they independently sense and stochastically interact with other agents and/or their local environment in order to produce complex, system-level behaviour.

ABMs utilise computational power in order to explore the dynamics of complex systems that can be difficult to solve by purely analytic techniques or other approximate mathematical methods. In order to simulate complex adaptive systems, for instance, an agent's behaviour may include capabilities to evolve in order that newer behaviour may emerge. Learning and adaptation protocols such as neural networks, evolutionary

algorithms, etc., can allow for such evolving agents [50, 51, 52, 53]. A typical ABM has three components:

- A set of autonomous agents who are self-contained discrete individual with attributes, behaviours and decision-making capabilities. Agents are also social (they interact with other agents) and have resource attributes and goals, and can modify their behaviour by adaptation.
- Agent relationships and methods of interaction of agents with each other and with the environment, and
- An environment that can have its own attributes and provide a spatial landscape for agents to interact in [53].

Given a sufficient level of detail that can be coded into the properties of each agent, ABM can be used in a wide range of situations, including as a descriptive tool and as a substitute to real world experiments. In a biological context, especially in ecology, ABMs have been used to predicting the behaviour of populations given some known properties of individuals [54]. In a similar way, given information about a population, one can attempt to determine the behaviour of individuals that is necessary to achieve the same [21]. Most useful applications have been made in order to see the response of a given system to a change in the environment - for instance, given details of individual behaviour, what effect does a change of the environment have on population indicators. Such studies have been of great utility in modeling fish population dynamics [55].

1.5 Genome analysis

The genome of an organism is, roughly speaking, all its genetic material put together. Genomics - the study of the genome - has a wide horizon since in some sense all properties emanate from the genome. Thus genome structure, function, inheritance, evolution, all aspects are important to understand. The first steps toward understanding a genome is its sequencing, assembly and annotation [56, 57, 58]. At the present time, there are a number of tools for genome sequencing, some of which have made the process routine [59]. Once the sequence, namely the order of nucleotides in a DNA element is known, the entire genome can be reconstructed via assembly. Genome annotation is the process of determining the role of each given stretch of the genome: whether it corresponds to a gene, to a regulatory sequence, or corresponds to other functional elements. In short, this gives meaning to the sequenced genome.

Annotation needs to be done at the nucleotide level, but it is also important to determine gene expression and biological function [56]. Nucleotide level annotation maps

various genetic elements such as coding and non-coding regions, regulatory regions, genes and ORFs, tRNAs, rRNAs, SNPs etc.: this is largely done by comparison with known sequences [60]. Expression level annotation requires the cataloguing of proteins, assigning putative functions and so on. At the process level annotation implies the association of genes to biological processes like metabolism, cell division, cell death, photosynthesis, cellular transport etc., Many recent advances have led to high-throughput genome sequencing and assembly. Computational bioinformatics techniques have been developed in order to annotate genomes automatically and this information can be used for functional, meta- and comparative analysis of genomes [58].

1.5.1 COGs: Cluster of Orthologous Groups

The subject of genomics is vast and is described in several comprehensive texts [57, 60, 61]. Here I discuss the basic idea behind the formation of COGs, namely the classification of genes into clusters of orthologous groups that was introduced by Tatusov, Koonin, and Lipman [62].

Many genomes have only partial experimental functional verification of their putative annotation. For comparing genomes, though, such information is essential. Since on average, functionally important genes are highly conserved evolutionarily, a classification of genes into clusters of orthologous groups can be used for computational functional annotation of genes [62]. Orthologs are genes that share an ancestor by vertical descent, while genes within the same genome that are related by duplication are termed paralogs. COGs are constructed by comparing protein sequences from complete genomes. When comparing distinct genomes, if a group of proteins are more similar to each other than to other proteins of their own genomes, they are likely to belong to the same orthologous family and so can be constituted as a COG [63]. By homology, the functional information of characterized proteins can be imparted to its homologue which is uncharacterised, thereby COGs can be used to functionally annotate unknown sequences and genomes.

There are at present a large number of repositories which contain annotated information of many genomes and for the analysis presented in Chapter 5 the NCBI genome database [64] and IMG (the Integrated Microbial Genomes) database [65] were used.

1.6 Branching process

In modelling cell reproduction or viral reproduction, it is useful to consider these as branching processes. For example, in course of development of cell types, a multipotent cell produces other multioptent cells, but depending on the physiological state it can also produce other specific types of cells. This creates an hierarchy: differentiated cells

proliferate inside a multipotent cell, which themselves further proliferate. Alternately, consider bacteriophage inside a bacterium, both of which replicate. When the bacterium divides, at the next generation, there can be different numbers of phage replicants in each of the daughters.

A branching process is a convenient mathematical and computational technique that can be used to describe the development of populations. A branching process model consists of entities that survive for a given period of time (which can be random) and produce a given number of progeny (which may also be random). Such populations are said to have the branching property, namely the assumption that the entities or particles in the process are self-recurrent [66]. In a classical branching process (see Fig. 1.9 (a)) the progeny are produced at the point when the parent ceases to exist. If progeny are produced during the lifetime of the parent (Fig. 1.9 (b)), it is known as general branching process [67]. The classical process can be used to model single cell populations, molecular processes like DNA replication and gene evolution, while the general process is used to model populations of multicellular organisms like plants and animals. In a continuous-time Markov models of a branching process, the classical process can be redefined as a general process by assuming that at the point of reproduction, one of the progeny can be assumed as an extension of the parent [66]. Different types of branching processes can be envisioned, depending on the distribution of particle lifetimes, the type-space (a set of all types of particles), and on the average number of progeny, μ. For Galton-watson [68] process the time is a discrete integer series and the random τ results in a age-dependent Bellman-Harris process [69]. A $\mu > 1$ leads to a super critical process, a $\mu < 1$ leads to subcritical process and a $\mu = 1$ leads to a critical process.

Branching processes are conceptually simple and can model coarse-grained biological systems such as the Galton-Watson process [68] for population growth, or even very detailed processes (using markov-multitype branching process). A fair level of detail can be incorporated so that branching process models can provide quantitative predictions and be used in studies of molecular and cell biology [70, 71, 72], immunology [73, 74, 75, 76], demography [77], viral infections [78], genetics and evolutionary theory [79].

1.7 Machine Learning Techniques

In a subsequent chapter of this thesis, I have used machine learning, namely an automated method for detecting meaningful patterns from data without assuming an equation as a model [80]. Common to most machine learning techniques is a training set from which patterns are learned via a specified algorithm, and this can then be used to act on or predict the outcome of other data, the test set. Formally, a "computer program is said to

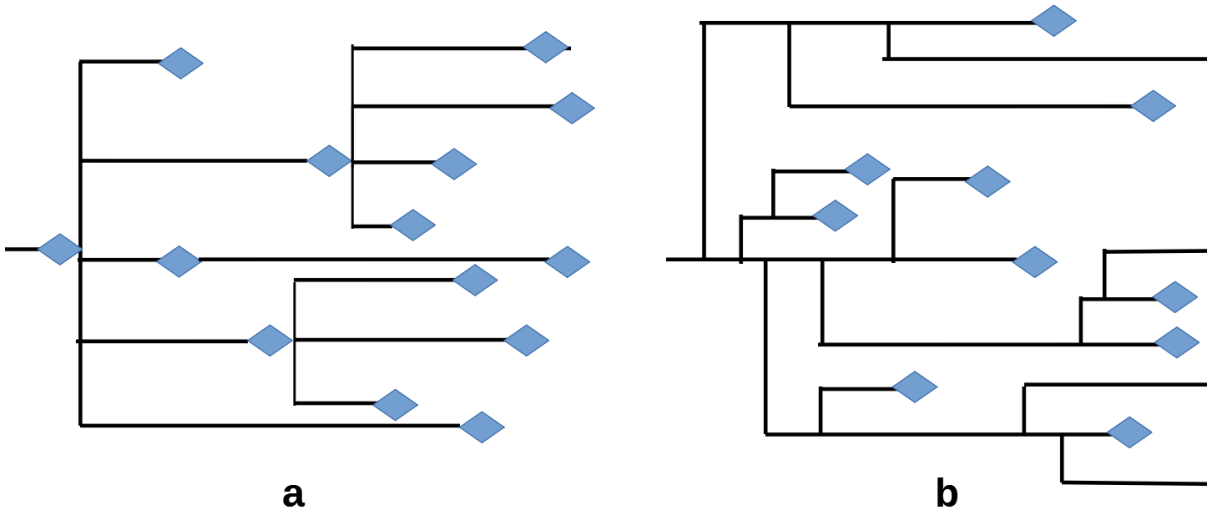

Figure 1.9: Representation of branching processes as trees, the solid line represents the lifetime of the entity, and the diamond represents the point of "death". The branches denote birth events. Time is measured along the x-axis. (a) Branching process with progeny produced at death, and (b) branching process with progeny produced during the lifetime of the parent.

learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [81].

With the increase in computational resources, the usage of machine learning had tremendously increased in past few decades and is spread across all fields of learning and technology. Machine learning is routinely used in search optimisation, spam filtering, credit card fraud detection, phylogenetic analysis, disease detection, tumour detection, recovery rates in patients, drug discovery, DNA sequencing, natural language processing, face detection etc., [80, 81].

Supervised and unsupervised learning are two broad categories of machine learning techniques [81]. In the latter, the training data is unlabelled and the task is to decipher its intrinsic structure. This process of grouping objects such that objects in group are more similar to each other than to others is termed clustering. In supervised learning, the training data is labelled and the task of machine learning is to assign labels to the test data. Classification and regression are common examples of supervised learning, and a large number of techniques have been developed over the years to examine data for these

purposes. A number of clustering methods are known and include the K-means [82, 83], K-medoids[84], hierarchical clustering [85], neural networks [86], hidden Markov models [87]. Other classifiers include support vector machines [88], naive Bayes algorithms [89], etc. (See Fig. 1.10). Below I describe two learning methods, decision trees and the random forest model, that are used in a subsequent Chapter.

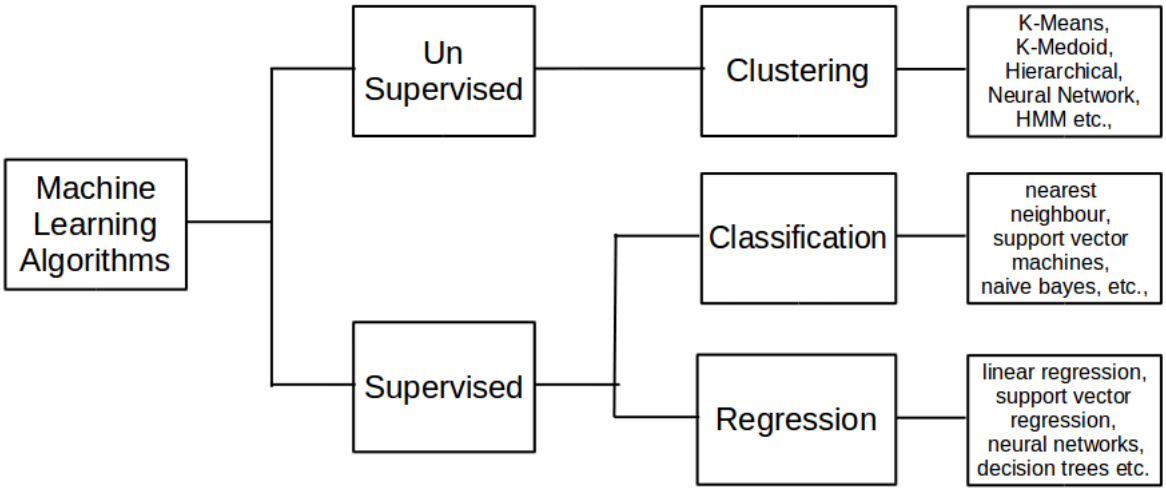

Figure 1.10: Classification of Machine Learning Algorithms and examples.

1.7.1 Decision trees

The so-called decision tree is a graph (with a tree structure) on which all possible outcomes can be depicted as a series of sequential decisions [90]. There is a root node (with the entire population under consideration) and internal nodes which split to sub-nodes. Each branch is an outcome of the decision at a decision node. A terminal-node (or leaf) does not further divide [91, 92].

A decision tree for protein-protein interactions is shown in Fig. 1.11. A is the root node, $B-E$ are decision nodes and $F-J$ are leaves. In the simplest case, the decisions have binary outcomes: each internal node has yes/no sub-nodes. By asking a series of questions about the features by following the path from the root node onwards, an object can be classified into associated labels. Other decision trees can have more than two branches, leaves might denote a probability distribution, etc. Generating such decision trees from a training set and using them for classifying a test set, is decision tree learning. If all possible decision trees can be built for training data, the best can be chosen for classifying test data, but in most cases that is computationally inhibiting. Hence various algorithms are used for generation of decision trees from training data. Some of them are ID3 (Iterative Dichotomiser 3), CART (Classication And Regression Tree), CHAID

(CHi-squared Automatic Interaction Detector), MARS, etc., For choosing the decision nodes these algorithms use various metrics like gini impurity, information gain (entropy), chi-squared error, variation reduction etc., [91].

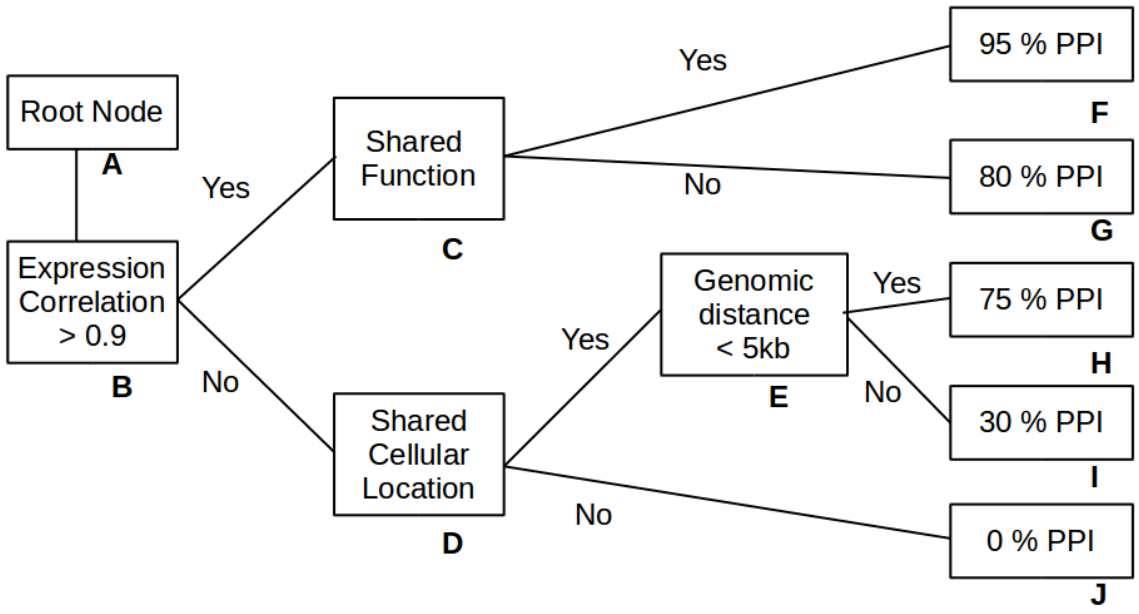

Figure 1.11: Example of a decision tree that can be employed for identification of protein-protein Interactions.

1.7.2 Random Forest method

In many cases of classification, decision trees give fairly high accuracy. Further increase in accuracy can be achieved by aggregating the results of an ensemble of decision trees using strategies such as boosting [93] and the random forest method [94].

In the random forest method, using a randomised algorithm n distinct decision trees are grown. Each unpruned classification or regression tree is grown using a bootstrap sample from original data. To grow such trees at each node by choosing the best split, a random subset of m features is considered [94, 95]. Thus the number of trees in the forest and the number of features considered at each split, n and m respectively, are the only two parameters for a random forest. Typically, the model is not very sensitive to these parameters. The random forest strategy performs well in comparison to other classifiers such as support vector machines and neural networks [94] and is robust against overfitting.

1.8 Networks

The network paradigm is currently of great interest in a wide variety of areas. Extensive applications have been made over the past decades to address a number of biological problems ranging from metabolic networks, to genetic interactions, cellular and metabolic pathways, regulatory and interaction networks in other areas such as food webs, the World-Wide Web, airline networks, scientific collaborations and so on.

Since this is a topic that has been covered extensively in textbooks [96, 97], the discussion on the essential features of networks will be brief. A network is essentially a graph, namely a mathematical structure used as model for pairwise relations between objects, that consists of nodes or vertices and edge or links. Shown in Fig. 1.12) (a) are the set of vertices $V = a, b, c, d$ that have edges that can be enumerated $E = \{(a, b), (a, c), (b, c), (c, d)\}$, the vertices being represented by dots and the edges by lines. The adjacency matrix $\mathcal{A}$ of a network of n nodes is a $n \times n$ matrix, with elements a_{ij} that are zero if nodes i and j do not have a link, and nonzero if there is a link. The adjacency list $\mathcal{E}$ enumerates all the linked nodes.

1.8.1 Network topologies and characterization

Networks can be classified as directed ($a_{ij} \neq a_{ji}$) or undirected ($a_{ij} = a_{ji}$) (Fig. 1.12 (a), (d)), and furthermore the edges can have weights associated with them. Trees are characterised as having no loops, while multigraphs have more than one edge connecting the same vertices (Fig. 1.12 (e)).

In a k-regular graph (Fig. 1.13(a) is 3-regular graph) every node has k connections. In a complete graph, on the other hand, all possible pair of vertices are connected by edges (Fig. 1.13 (a)), and if there is at least one pair of vertices which are not connected by a path then the graph is disconnected; clearly it can be split into a number of connected components (Fig. 1.13 (b)). In a bipartite graph, vertices can be grouped into two sets such that no two vertices inside a each group are connected by an edge (Fig. 1.13 (c)).

In a cyclic graph if the edges are ordered as $V_1, V_2, ...V_n$ then the edges are $\{v_i, v_i + 1\}$ where $i = 1, 2, ..., n - 1$, and the edge $\{v_n, v_1\}$. A graph with no cycles is a tree, and a planar graph is one that can be represented on a plane without any crossing over between edges (Fig. 1.12 (a,d,f) and Fig. 1.13 (b,d)).

The size of network is the total number of vertices, and the degree of a node (d_i) is the total number of edges attached to it. For an undirected graph, the degree of a node

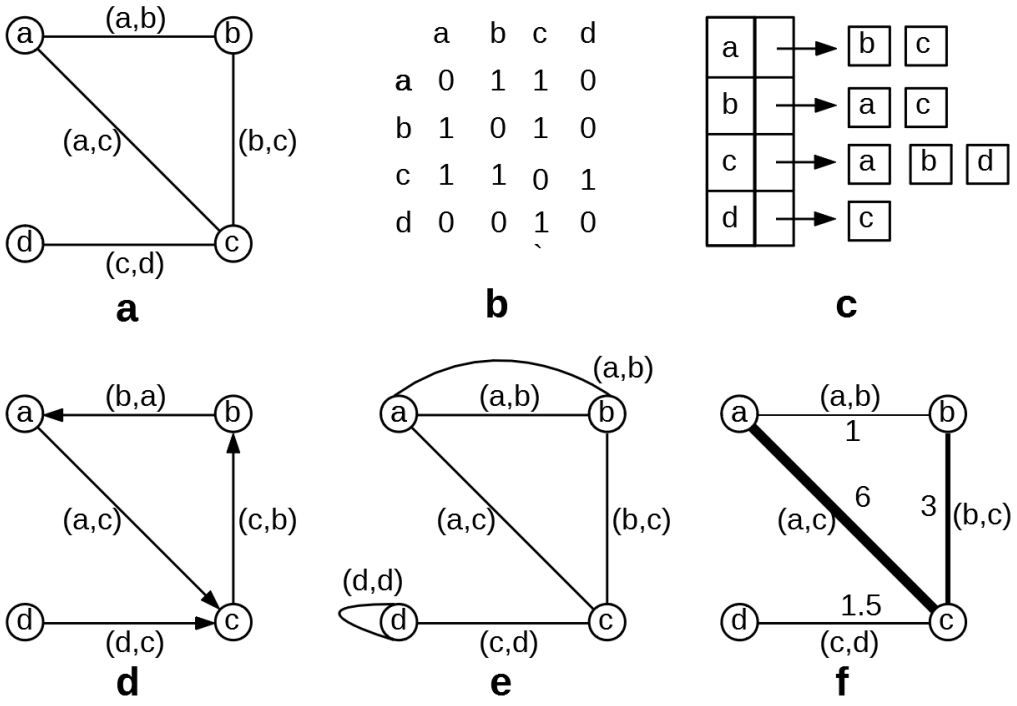

Figure 1.12: Schematic representations of types of graphs. (a) Undirected simple graph, d) directed graph, e) multigraph with a loop and multiedges, f) weighted graph.

is related to the entries in the adjacency matrix, $\mathcal{A}$,

$$d_i = \sum_{j=1}^{n} a_{ij}.$$

A directed graph has an in-degree, namely the number of incoming edges and the out-degree, the number of outgoing edges. A trail is the sequences of edges, where all edges are distinct. A path is a trail with all distinct vertices (except for a loop where starting and last vertices are the same). The length of a path is the total number of edges for a unweighed graph or weighted sum for a weighted graph. A k length walk is denoted by the alternating sequence of nodes and edges (v_0, e_0, v_1, e_1,...., v_{k-1}, e_{k-1}, v_k). The distance between two vertices is the shortest path between the two. Nodes with high degree are hubs, but the relative importance of a vertex can be quantified through various types of so–called centrality measures. The closeness centrality of a node m is the average length of shortest paths to all other nodes,

$$C_c(m) = \frac{1}{n-1} \sum_{j \neq i} d_{ij},$$

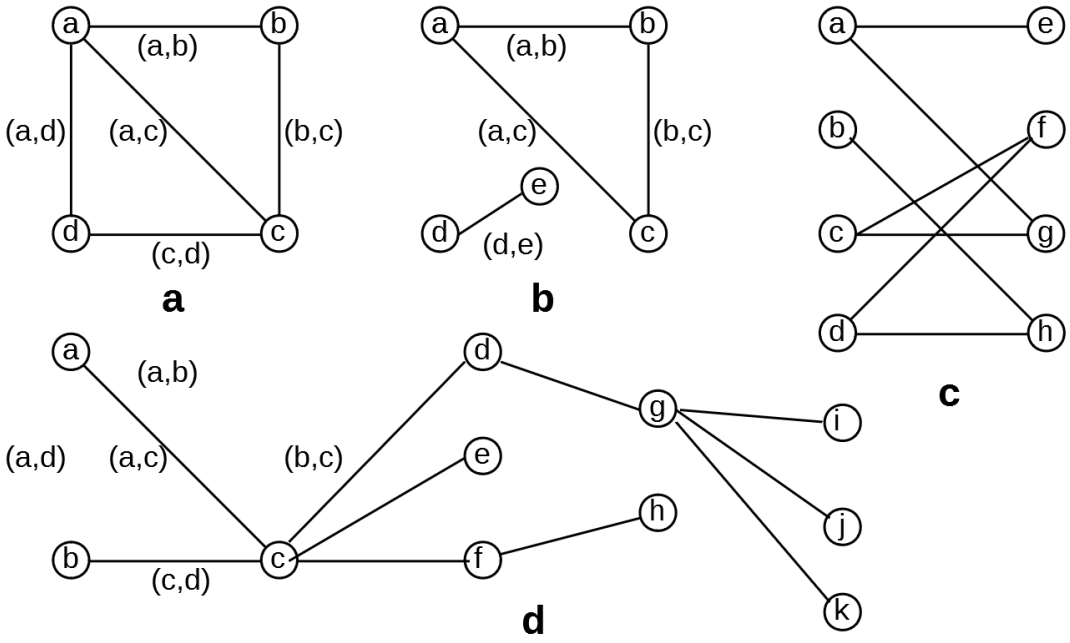

Figure 1.13: Types of graphs: a) 3-regular and complete graph, b) disconnected graph with two components , c) bipartite graph, d) tree graph with no cycles.

where d_{ij} is the distance between nodes i and j, and the distance between unconnected nodes is n. The closeness of a node lies between 0 and 1. Betweenness centrality is the sum of fraction of times a node falls on the shortest path/s between pairs of all other nodes. The betweenness centrality of node m is

$$C_b(m) = \sum_{m \neq i \neq j} \frac{\sigma_{ij}(m)}{\sigma_{ij}},$$

where $\sigma_{ij}(m)$ is the number of times shortest path between i and j includes m and σ_{ij} is the total number of shortest paths between i and j. The distribution of degrees is an important metric of a network, many real-world networks have a power-law degree distribution and are termed scale-free. The diameter of a network is the largest distance (path) in the network. A triplet in a network is a group of three connected nodes. If all the nodes in a triplet are connected it is a closed triplet (triangle) else is an open triplet. The clustering coefficient of a network is defined as three times the ratio of all closed triplets to all connected triplets ($C = N_\Delta / N_\wedge$).

In a scale-free network, due to the power law dependence, most nodes have few neighbours while a few have many neighbours. Thus, although most pairs of nodes are not neighbours, given a pair of unconnected nodes, they are likely have common

neighbours, and the shortest paths are fairly small: this is sometimes termed the small-world property [98]. Such networks have a low value of the shortest path and relatively large clustering coefficients as compared to networks generated by random processes.

1.9 Plan of the thesis

The individuals of many species interact among themselves and a set of all such interactions define their social behaviour [99]. Some insect species such as termites, bees, and wasps, show a high form of social behaviour termed *eusocilaity*. In an eusocial population, the individuals are organised in colonies or nests and are segregated into workers and queens. A queen is a reproductive female, while a worker is non-reproductive. Both may cooperatively care for the progeny [100, 101, 102] in the colony in which overlapping generations will coexist. An associated feature of eusociality is size dimorphism: queens are significantly larger than the workers, but there is an associated asymmetry: although queens are a few times larger than workers, the lifespan of queens is several orders of magnitude larger than that of workers [104]. This is a very drastic departure from Kleiber's law [105], and considering the size of the queen, such a long lifespan is an anomaly. In Chapter 2 I discuss the theories of evolution of ageing in this context and computationally explore the possibility of long lifespans of queen being an outcome (or collective behaviour) of the population structure in eusocial colonies.

I extended the multilevel evolutionary model of eusociality by Nowak, Tarnita, and Wilson [49] to include age structure. I also defined an agent based model of the same and extended it for polygynous species. By comparing the fitness of each of these populations, as observed in nature, I showed that the solitary species tend to have shorter lifespans while the monogynous or polygynous eusocial species are associated with longer lifespans. It is observed that in long-life populations, parameter range corresponding to the evolution of eusociality increases in comparison to the age-independent models. I argue that the evolution of long lifespans is thus intrinsic to eusociality, and is both a product as well as an enabler of eusociality.

Mammalian species can display a wide range of social behaviours, from the near solitary behaviour of orang utans, to the communal organization of several dozens of individuals in primate colonies, to thousands of individuals in bat colonies. Vampire bats show long term social behaviour such as social grooming and food sharing [106]. Food sharing provides the recipient a direct benefit, but when reciprocity is established such behaviour can increase the fitness of whole population. The social behaviours can also provide by-product (non-direct) benefits to the population. In Chapter 3 I model the non-kin resources sharing in vampire bats to quantify such by-product group ben-

efits. I used an agent based model (ABM) to simulate the blood sharing behaviour of vampire bats and compared group parameters between sharing and non-sharing populations. For constant ecological resources, I observe that, when compared to a non sharing population, resources sharing behaviour had increased the sustainable population size by 50%, increased the total resource accumulated by 10%, and reduced the average resource required per individual by 36%. I also show that the increase in cooperativity has a nonlinear effect on group benefits. Substantial group benefits are shown only after cooperativity is 60% and it increases exponentially to a maximum thereafter.

At a different level of biological organisation, the cells communicate with each other, with the environment and temporally respond to the sensible external cues. The process of such communication is termed as signal transduction or cell signalling [107] and is crucial to governing and coordinating cellular behaviour. During important biological functions such as homeostasis, development, immune response, tissue repair etc. [107], cells respond to the environment using signal transduction and behave as an ensemble. NFκB an important protein complex nuclear factor, is found in almost all types of animal cells and regulates the DNA transcription [108]. It is shown to be a constituent of many signal transduction pathways and thus participate in the cellular responses to the external stimuli.

Oscillations are ubiquitous in biology, and the dynamics of the transcription factor nuclear factor kB (NFκB) has been established experimentally to have oscillations, and several mathematical models have been developed as well. Given that the signal transduction pathways are generally coupled and can have common environment, the cells act as an ensemble. In Chapter 4 I study the collective dynamics of such cellular ensembles. This chapter is dedicated to understand emergence of interesting collective states that are possible in such coupled oscillators, such as synchronization, splay states, cluster synchrony and chimeras. Two types of ensembles are defined, one coupled by external TNF oscillator and other by global mean filed coupling of NfκB molecule. I observe all the above defined states in these ensembles.

In multi-stable regions of NfκB dynamics, the behaviour of basins of attraction plays an important role in understanding the effect of noise, which is intrinsic to such cellular systems. I show that the basin boundary has a fractal dimension and measure the uncertainty coefficient of basins, which indicate the difficulty to choose the final state given a range of initial conditions. This show that increase in the amplitude of external TNF oscillator increases the uncertainty coefficient and its period reduces the uncertainty coefficient.

At the evolutionary scales of segregation of living organisms into prokaryotes and eukaryotes, the comparative studies show that each group have many specific character-

istics. Understanding the evolution of such group specific behaviours is an active research area. In Chapter 5 I discuss one such feature, that of genome size. I compared the distributions of number of genes and genome lengths across essentially all available prokaryotic, archaeal and eukaryotic organisms. There is a sharp separation between prokaryotes and eukaryotes, with prokaryotes limiting their gene length to 10 Mbp and gene numbers to ten thousand, while eukaryotes have exceeded these limits. Currently extant bacteria and eukaryotes diverged after LECA, the last eukaryotic common ancestor.

To comparatively understand the mechanism of such a constraint, I analysed the enrichment of various COGs (clusters of orthologous genes) with genome size both in prokaryotes and eukaryotes. As has been observed in earlier studies as well, the power law scaling of transcriptional regulation genes with an exponent greater than 1 is at the crux of constraints on genome size in prokaryotes. I also calculated and compared the subcellular energetics of prokaryotes and eukaryotes and the difference between prokaryotes and eukaryotes is just by a factor of 1 to 8. The mechanism through which eukaryotes bypassed such a limit has to be determined. The scaling laws for genes involved in transcription and the energy related genes of prokaryotes suggest that the opportunity cost, namely the number of transcription genes per energy related genes increases with the genome size. Genomic asymmetry between LECA and the protomitochondrion during eukaryogenesis will result in a asymmetric opportunity costs for energy related protein production. I hypothesise that this asymmetric opportunity cost could have led to enrichment of energy genes in protomitochondrion, and could have supplied the necessary increase of protein production for genome expansion of LECA. The increased expression by increase in genome copies of protomitochondrion would lead to over expression of many genes. To reduce such a cost, large parts of its genome are transferred to the nuclear genome and thus helps in its remodelling, and in this sense, one may infer that the eukaryotic cell could have used mitochondrial colonies in order to increase the size and complexity of the nuclear genome.

Two ongoing research problems model cellular populations and generate a interspecies molecular interaction network. I discuss the preliminary results of these works is Chapter 6. The branching process model of CD8+ T cell, replicates the experimentally observed heterogeneous differentiation patterns and I generated a partial PPI (protein-protein interaction) network for *Helicobacter pylori* and *Homo sapiens* using a random forest method. The PPI network is further analysed. The concluding chapter presents a summary of thesis and outlines future research directions.

CHAPTER 2

Modeling Long Lifespans in Eusocial Insect Populations

Along with division of labour, and life-history complexities, a characteristic of eusocial insect societies is the greatly extended lifespan for queens. The colony structure reduces the extrinsic mortality of the queen, and according to classical evolutionary theories of ageing, this greatly increases the lifespan. I explore the relationship between the evolution of longevity and the evolution of eusociality by introducing age-structure into a previously proposed evolutionary model and also define an associated agent-based model. A set of three population structures are defined: (i) solitary with all reproductive individuals, (ii) monogynous eusocial with a single queen, and (iii) polygynous eusocial, with multiple queens.

In order to compare the relative fitnesses I compete all possible pairs of these strategies as well as all three together, analysing the effects of parameters such as the probability of progeny migration, group benefits, and extrinsic mortality on the evolution of long lifespans. Simulations suggest that long lifespans appear to evolve only in eusocial populations, and further, that long lifespans enlarge the region of parameter space where eusociality evolves. When all three population strategies compete, the agent-based simulations indicate that solitary strategies are largely confined to shorter lifespans. For long lifespan strategies the solitary behaviour results only for extreme (very low or very high) migration probability. For median and small values of migration probability, the polygynous eusocial and monogynous eusocial strategies give advantage to the population respectively. For a given migration probability, with an increase in lifespan, the dominant strategy changes from solitary to polygynous to monogynous eusociality. The evolution of a long lifespan is thus closely linked to the evolution of eusociality, and our results are

in accord with the observation that the breeding female in monogynous eusocial species has a longer lifespan than those in solitary or polygynous eusocial species.

2.1 Introduction

For living organisms death results from both extrinsic causes such as predation, disease, or accident, and intrinsic causes that include senescence. The force of natural selection decreases with age, with genes that affect later life being under reduced selection pressure [104, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118]. As a consequence, according to mutation accumulation theory, alleles with deleterious effects — mutations — accumulate at later ages leading to senescence [110, 111, 112, 116, 117, 118]. An alternate view put forth in the theory of antagonistic pleiotropy [111, 118, 119] is that senescence evolves due to side-effects that are deleterious when occurring late in life history, but which are favourable early on in the life cycle. These hypotheses propose that ageing can be evolved [115, 120, 121] and further, that an increase (or decrease) in extrinsic death-rate will increase (or decrease) the rate of ageing [109, 118, 122]. There are, of course, additional and more complex effects that are density and life-history dependent [109]. Some recent empirical and experimental observations give results contrary to simple classical predictions [123, 124] and the causes are largely unknown [125]. To explain the more complex effects and deviations, newer modeling approaches such as simulated annealing optimization [125], age-structured evolutionary models [126], hierarchical models [127] are being employed.

Eusociality, an advanced form of social organisation seen in several insect societies, is defined by reproducing (queen) and non-reproducing (worker) individuals that cooperatively care for the young [100, 101, 102]. Reproductive strategies depend on whether a colony can contain one or more queens: in monogynous species, each queen initiates a new and independent colony, while in polygynous species several queens remain in an established colony [103, 104, 128, 129]. Further, queens in monogynous eusocial societies have extraordinarily long lifespans as compared to those in polygynous species or compared to individuals of solitary species [104, 129, 130, 131]. The differences are striking: queens of the polygynous *Formica polyctena* and the monogynous *Formica exsecta*, both mound-building wood ants live 5 and 20 years respectively, while individuals of solitary species live at most for a month [104].

The interrelation between these two features has been studied previously [104, 127, 132] and indeed the occurrence of long lifespans in eusocial queens that also have a lower extrinsic death-rate is seen as a test of evolutionary theories [104]. Shorter lifespans that are associated with polygyny are understood within the framework of evolutionary

theory [104] under the assumption that polygynous species have a higher mortality risk of queens since they inhabit fragile nests that increase the extrinsic mortality. However, there are polygynous species that inhibit permanent nests with the lifespan of those queens being lower than a related monogynous queen [114], suggesting that the link between queen lifespan and eusociality needs further exploration. In this context to explore the interdependency in evolution of these features, newer evolutionary models are required. These models should allow for various competing strategies of both eusociality and ageing.

The Nowak, Tarnita, and Wilson [49] (NTW) model for the evolution of eusociality examines the population structure of eusocial organisms in order to understand the conditions for evolution of eusociality. The details of the model are as follows: The eusocial population consists of several colonies, each of which consists of a reproductive *queen* and non-reproductive workers. The total number of queens and workers in a colony is its size. Colonies of a particular size show similar behaviour (namely the death- and birth-rates). The queen and workers of a colony die at a characteristic death-rate while queens reproduce at a characteristic birth-rate (these rates are size dependent). Progeny either stay in the nest with probability q to become workers and increase the size of the parent colony by one, or migrate with probability (1-q) to start a new colony. Death of the queen kills the entire colony while death of a worker reduces the colony size by one (Fig 2.1.c). The model assumes that the benefits of a colony are realised only after the size of the colony reaches a threshold above which it stays constant. Solitary populations, on the other hand, consist of homogenous individuals which can reproduce and die at characteristic rates (Fig 2.1a). NTW compete the two populations and examine conditions when natural selection favours the eusocial allele. When large benefits are associated with colonies above a particular size, eusociality can evolve for a range of q by increasing the rate of oviposition and reducing the death-rate for queens. Ageing is absent in the model, the rate of oviposition and/or death-rate being age-invariant. There has been some debate on the conclusions that can be drawn from the initial studies [133].

Does the eusocial population structure leads to the evolution of long lifespans in queens? Keeping this underlying question in mind, the aim of the present work is to test the conditions for evolution of long lifespans in such eusocial populations. The NTW model operates within the standard theory of natural selection, making it possible to evaluate multiple competing hypotheses [49, 133]. I extended the NTW framework to build two models, one which introduces age-dependence in the reproductivity and one that uses agent based modeling to allow for *polygynous eusocial* life-history, in addition to solitary and monogynous eusociality. I use both models to determine how the population structure, and specific parameters affect the evolution of long lifespans, and how this can

in turn affect the evolution of eusociality.

2.2 Model

I first briefly review the essential components of the models studied by NTW. For a solitary population model the abundance of solitary individuals, denoted x, evolves in time according to the dynamics

$$\frac{dx}{dt} \equiv \dot{x} = f_s(x)$$

where the t is time and the function f_s contains details of the assumed models of birth and death. In the simplest case of constant birth-rate b and death-rate d, $f_s(x) = (b-d)x$ (Fig 2.1a).

In an eusocial population model, one considers x_j colonies (or nests) of size j with j = 1,2,. . . , n. Where n is the largest nest size. The total population is

$$P = \sum_{j=1}^{n} j x_j.$$

The population of a nest changes by growth and by migration, namely

$$\dot{x}_j = b_{j-1} q x_{j-1} - (b_j q + d_j) x_j$$

with $j = 2, 3, \ldots$, while for nests of size 1 the dynamics is

$$\dot{x}_1 = \sum_{j=1}^{n} b_j (1-q) x_j - (b_1 q + d_1) x_1$$

where b_j and d_j are the birth and death rates for colonies of size j. The probability that progeny migrate from the parent colony is $1 - q$; see Fig 2.1.c.

Features such as density limitations and worker mortality can be included in the model and NTW have extensively studied these. Our present interest is in the inclusion of age-dependence within this general framework. I discuss this variation next, first within the NTW model, and thereafter in an agent-based model that has a greater level of flexibility in implementing the population structure. Parameters and variables used in these models are summarised in Table 2.1.

2.2.1 Age-dependence in evolutionary dynamics models

To understand the effect of age-structure on a solitary population I define a continuous population model [11]. It is important to include the effect of age a in order to incorporate

Table 2.1: **Summary of the model variables and parameters.**

Var.	Description	Comment
x	Population size	Total population for solitary NTW model.
b	Birth-rate	Rate of progeny birth for solitary NTW model.
d	Death-rate	Rate of death in solitary NTW model.
n	Largest nest size	Size of the largest nest in eusocial models. Choosen as 30.
x_j	Size of size-class	Number of colonies of size j in eusocial NTW model.
P	Population size	Total population for eusocial NTW model.
b_j	Birth-rate	Rate of progeny birth for colony of size j in an eusocial NTW model.
d_j	Death-rate	Rate of death of queens for colony of size j in an eusocial NTW model.
q	Stay put probability of progeny	In eusocial populations, it is the rate of progeny to stay in the parent colony to become a worker.
β	Effective birth-rate	Rate of progeny birth for a continuous model (effective birth-rate when the intrinsic death-rate is subtracted).
μ	Extr. deathrate	Rate of extrinsic death in continuous age-structured population model.
S	Population size	Total population for continuous age-structured model.
m	Max age	It is the last age in finite age population models, choosen as 30.
r	Max rep. age	The age or age-group after which birth-rate is equal to zero.
K	Lifetime reproductive capacity	Total progeny produced in its lifetime. A constant, assuming lifetime reproductive investment is same across comparing models.
h_1	Age 1 birth-rate	The starting age or age-group birth-rate.
h_2	Birth-rate at age group r	In finite age population models, this is the birth-rate at last reproductive age or age-group r.
b_j^i	Discrete model Birth-rate	Rate of progeny birth for age-structured NTW model. Depends on age (i) and size (j) of the colony.
d_j^i	Extrinsic death-rate	Rate of extrinsic death of queens in age-structured NTW model. Low enough to avoid extinction. Depends on age (i) and size (j).
g	Group (colony) benefit	Defines the benefits due to colony structure. In our simulations this takes values in the interval [1,10].
x_j^i	Size of an age-size class	Number of nests in particular age (i)-size (j) class. It denotes the number of colonies of size j and with queen of age i.
ψ	Time progression rate	Rate of transfer of colonies to the next age class. Chosen as 0.25, does-not affect results qualitatively.
α_j	Colony size reduction rate	Product of worker death rate and number of workers in the colony. It defines the rate at which colonies move to a lower size-class.
ϕ	Density dependence factor	In age-structured NTW models, it is defined as $1/(1+U)$, where U is the total population and can be equal to x_E or x_S or $x_E + x_S$.
x_S	Population size (fitness)	Total population of age-structured solitary NTW model and is the absolute fitness of that population.
x_E	Population size (fitness)	Total population of age-structured eusocial NTW model and is the absolute fitness of that population.
p	Age at maximum birth-rate	In a unimodal (tent-shaped) reproductive strategy, the birth-rate has a maximum at an intermediate age.

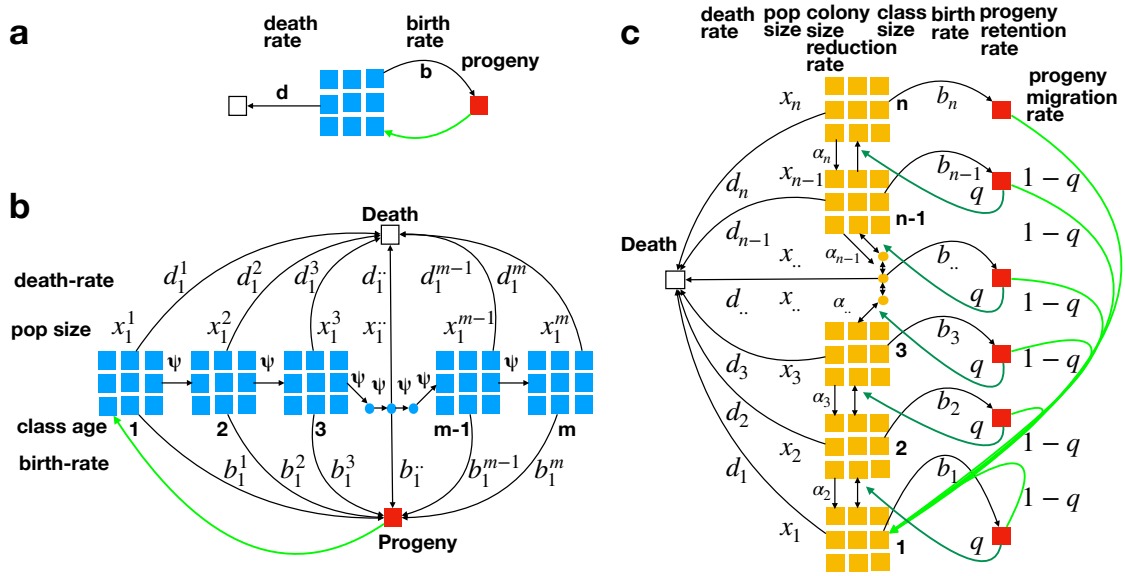

Figure 2.1: Graphical depictions of various evolutionary models. A single reproductive individual is represented by a blue square, the eusocial colony by an orange square, and red square represent progeny. Subscripts on variables denotes the size of colony, and superscripts denote the age of the queen. a) Solitary model: individuals reproduce at rate b, die at rate d, and the progeny joins the population. b) Age structured solitary model: the populations is segregated into age-classes, each of which move to next age-class at rate ψ. All individuals reproduce at rate b_1^i and die at rate d_1^i, and the progeny joins the first age class. c) Monogynous eusocial model (NTW): The eusocial colonies are segregated into size-classes and the queen of each colony reproduces at rate b_j^i and die at rate d_j^i. The progeny can remain to increase the size of the parent colony by 1 or start a new colony of size 1. Workers in a colony reduce the size by 1 upon death.

different breeding patterns, and I first obtain an expression that defines the fitness of a population in terms of the ageing strategy. For a population with finite age structure, the critical threshold S (or fitness) for growth is defined as [11]

$$S = \int_0^m \beta(a) e^{-\int_0^a \mu(v)dv} da, \tag{2.1}$$

where $\beta(a)$ is an age-dependent effective birth-rate, m is the maximum age and the rate of death due to extrinsic causes is μ. Assuming that this is an age-independent constant D gives

$$S = \int_0^m \beta(a) e^{-aD} da. \tag{2.2}$$

The functional dependence of the effective birth-rate, namely $\beta(a)$ defines the "ageing

strategy" of the population. I consider $\beta(a)$ to be linear, interpolating from an initial rate of $\beta(0) = h_1$ to a final $\beta(r) = h_2$ with $\beta(a) = 0$ beyond, namely for $a > r$ (Fig. 2.2a). I chose this function and later a tent function (with an additional parameter p) to define ageing so to approximate the unimodal ageing behaviour seen in nature, where the effective birth-rate starts at $\beta(0) = h_1$ first increases till age p and then decreases by age r ($\beta(r) = h_2$) (Fig. 2.2b). Choices of h_1, h_2 and r will specify the ageing strategy (Fig. 2.2a) in the following way. For any form of $\beta(a)$, the total reproductivity is kept constant at K [134]. Thus

$$K = \int_0^m \beta(a)da \tag{2.3}$$

For the case above, namely $\beta(a) = 0$ for $a > r$ and linear in $[0,\ r]$ with $\beta(0) = h_1$, $\beta(r) = h_2$, I get

$$K = r(h_1 + \frac{h_2 - h_1}{2}),$$

which gives

$$h_2 = \frac{2K}{r} - h_1.$$

Thus for fixed K and r, either h_1 or h_2 can be chosen. The optimal strategy, namely the maximal S for a constant life-time reproductive capacity K defines a variational problem (within the linear choice for β). Solving Eq. (2) using this condition yields

$$S = \frac{h_2 r\,(2 - Dr) + 2K\,(Dr - 1) + e^{-Dr}\,(2K - h_2 r\,(Dr + 2))}{D^2 r^2}. \tag{2.4}$$

For given K, r and D, S will reduce with increasing h_2: the optimal strategy is to have higher early effective birth-rate. A population of solitary individuals thus cannot achieve a long-life life history. As effective birth-rate is defined as the difference between birth-rate and intrinsic death-rate, the ageing strategy can be achieved by various combinations of birth-rate and death-rate functions.

I include ageing structure within the NTW model for the evolutionary dynamics of solitary and eusocial organisms [49] as follows.

- The age structured NTW solitary population is segregated into classes identified by age and size. Here x^i_j is the number of groups of age class $i = 1, 2, \ldots$ and size j, with j=1 being used to denote a solitary population. The ageing strategy of the population will define the birth (b^i_1) rates of each age class. Here and in the rest of this paper, the effective birth-rate b^i_1 has had the intrinsic death-rate subtracted. The extrinsic death-rate is denoted as d^i_1. Individuals in an age class move to next

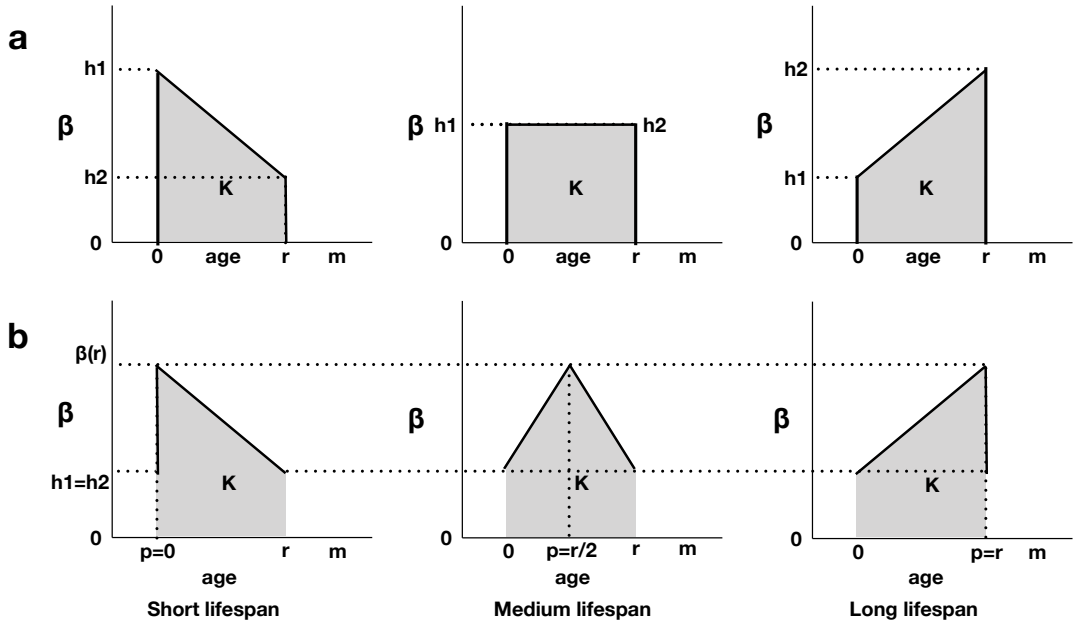

Figure 2.2: Graphical depictions of ageing-strategies. β denotes the effective birth-rate a) Simple monotonous ageing strategy and b) a unimodal (tent shaped) ageing strategy.

age class ($i \to i+1$) at ageing rate ψ, and progeny join the first age class (Fig 2.1b). For a population with m age classes, the master equation can be written as

$$\dot{x}_1^i = -(\psi + (1-\psi)d_1^i)x_1^i + \psi(1-d_1^{i-1})x_1^{i-1}, \quad i = 2, \ldots, m \tag{2.5}$$

$$\dot{x}_1^1 = -(\psi + (1-\psi)d_1^1)x_1^1 + \phi \sum_{i=1}^{m} b_1^i x_1^i \tag{2.6}$$

$\dot{x}_1^i$ being the change in population for each age class other than the first. The starting age class is defined by x_1^1 and $x_S = \sum_{i=1}^{m} x_1^i$ is the total population size.

- For an age structured NTW monogynous eusocial population, $j \geqslant 1$. There are both queens and workers, with nests being segregated into age-size classes. The properties of each eusocial colony (birth-rate, death-rate etc.,) are dependent on its age-size class. The population size of each age-size class (number of colonies) is denoted as x_j^i, where age is given by i and colony size (number of workers and queen) is given by j. A monogynous eusocial colony has following six behaviours, 1) queen dies at specific extrinsic death-rate (d_j^i) and the colony dies, 2) worker bee dies and colony size reduces by 1, 3) queen reproduces at birth-rate (b_j^i), 4) the progeny stay in the colony at q and increases the colony size by 1, 5) the progeny migrate at $1-q$ to start a new colony of size 1 and not changing the size of the

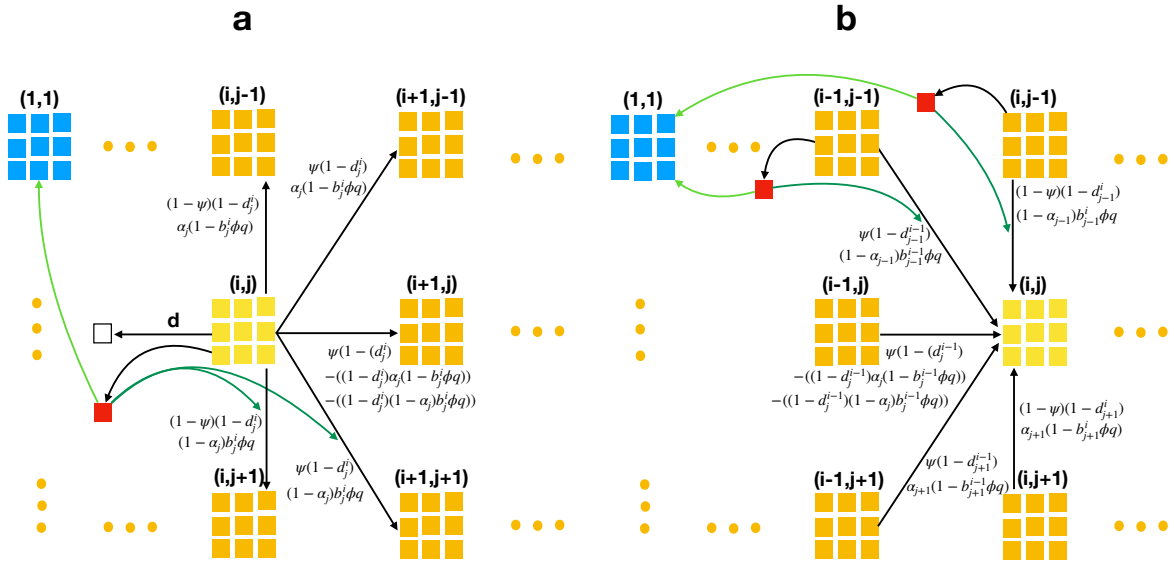

Figure 2.3: Graphical depiction of population dynamics in a monogynous eusocial model. A monogynous eusocial colony is indicated by an orange square. A group of yellow squares represents the age-size class whose dynamics are depicted. A blue square represents a colony of size 1, and a red squares represent progeny. The number of squares in a group is only for representation purposes and does not specify the size. The size of colony and the age class, namely the age of the queen is indicated above each group as an ordered pair (age,size). a) Depicts all possible dynamics for movement of colonies *out* of an age-size class, while b) illustrates all possible dynamics for movement of colonies *into* an age-size class.

parent colony, 6) the queen age at ageing-rate (ψ) and moves the colony to next age-class ($i \to i+1$). The combination of these behaviours define the dynamics of the population (Fig. 2.3).

The parameter α_j, the product of individual worker death-rate and number of workers in the colony defines the rate at which one age-size class transitions to another class with smaller size. If $d_{w,j}$ is the death-rate of workers in a colony of size j, then $\alpha_j = (j-1)d_{w,j}$. The appropriate master equation for a population of monogynous eusocial organisms is

$$\dot{x}^i_j = -d^i_j x^i_j + \sum_{i'j'} W_{ij \leftarrow i'j'} x^{i'}_{j'} - \sum_{i'j'} W_{i'j' \leftarrow ij} x^i_j \tag{2.7}$$

where the transition rates $W_{ij \leftarrow i'j'}$ indicate the rate of movement from the age-size group $i'j'$ to the age-size group ij. Fig 2.3 for description of $W_{ij \leftarrow i'j'}$ and $W_{i'j' \leftarrow ij}$:

$$
\begin{aligned}
W_{ij\leftarrow i'j'} &= \delta_{i,i'}\delta_{j,j'-1}(1-\psi)(1-d^i_{j+1})\alpha_{j+1}(1-b^i_{j+1}\phi q) \\
&\quad +\delta_{i,i'}\delta_{j,j'+1}(1-\psi)(1-d^i_{j-1})(1-\alpha_{j-1})b^i_{j-1}\phi q \\
&\quad +\delta_{i,i'+1}\delta_{j,j'-1}\psi(1-d^{i-1}_{j+1})\alpha_{j+1}(1-b^{i-1}_{j+1}\phi q) \\
&\quad +\delta_{i,i'+1}\delta_{j,j'+1}\psi(1-d^{i-1}_{j-1})(1-\alpha_{j-1})b^{i-1}_{j-1}\phi q \\
&\quad +\delta_{i,i'+1}\delta_{j,j'}\psi(1-(d^{i-1}_j)-((1-d^{i-1}_j)\alpha_j(1-b^{i-1}_j\phi q)) \\
&\quad\quad -((1-d^{i-1}_j)(1-\alpha_j)b^{i-1}_j\phi q)) \qquad (2.8) \\
W_{i'j'\leftarrow ij} &= \delta_{i,i'}\delta_{j,j'+1}(1-\psi)(1-d^i_j)\alpha_j(1-b^i_j\phi q) \\
&\quad +\delta_{i,i'}\delta_{j,j'-1}(1-\psi)(1-d^i_j)(1-\alpha_j)b^i_j\phi q \\
&\quad +\delta_{i,i'-1}\delta_{j,j'+1}\psi(1-d^i_j)\alpha_j(1-b^i_j\phi q) \\
&\quad +\delta_{i,i'-1}\delta_{j,j'-1}\psi(1-d^i_j)(1-\alpha_j)b^i_j\phi q \\
&\quad +\delta_{i,i'-1}\delta_{j,j'}\psi(1-(d^i_j)-((1-d^i_j)\alpha_j(1-b^i_j\phi q))-((1-d^i_j)(1-\alpha_j)b^i_j\phi q))
\end{aligned}
$$

and $\delta_{i,j}$ is the Kronecker delta namely 1 if $i=j$ and 0 if $i\neq j$. ϕ is the density limitation for growth of population, here taken to be $\phi = 1/(1+x_E)$ where x_E is the total population and is given by

$$x_E = \sum_{i=1}^{m}\sum_{j=1}^{n} j x^i_j \qquad (2.9)$$

Where m is the largest age-group and n is the maximum colony size. I keep the extrinsic death-rate constant and birth-rate is a linear function of age (Fig 2.2a). For colonies with size above a threshold (Th), the birth-rate is multiplied by the group benefit g, and the death-rate by the factor $1/g$.

Simulations were run separately for a population of eusocial organisms with a single queen (the monogynous case) as well as for a population of solitary organisms. In order to numerically solve the model, an age and colony size structured Leslie matrix L [135] of size $mn \times mn$ is constructed at each time-step t using the given transition rates.The population vector for for the eusocial model $\mathbf{X}_t \equiv (x^1_1, x^2_1..x^m_1, x^1_2, x^2_2....x^m_n)$ has mn elements. The population vector at time t+1 is given by

$$\mathbf{X}_{t+1} = L\mathbf{X}_t \qquad (2.10)$$

Given the population vector $\mathbf{X}$, the population size (x_E) at each time can be calculated using Eq. 2.9. For solitary models m=1 and the population measured is denoted x_S. The

above model is numerically solved using a standard R [136] code for a range of h_2 and D (the constant extrinsic death-rate) values until steady state is achieved. The code has been provided in the Supplementary Material.

In order to examine the effect of ageing on eusociality, I simulate both the monogynous eusocial model and the solitary models together, where the density limitation factor ϕ depends on the total population $x_E + x_S$ and is given by $\phi = 1/(x_E + x_S)$. Such competition experiments are conducted for a range of h_2, q and g values. When either x_E or x_S falls below a specified level, that particular strategy, namely eusociality or solitary, is deemed extinct.

2.2.2 Agent-based Models

A population can in principle implement a number of strategies: adopting a solitary lifestyle, eusociality with a single queen, or eusociality with several queens (polygyny). I need to model a more complex population structure, and agent-based modeling allows for an easier implementation of this complexity. I therefore study an agent-based model for age-structured evolutionary dynamics using RNETLOGO [137] and NETLOGO [138]. All codes are available in the Supplementary Material.

The agent-based description of the above evolutionary model is extended to include the polygynous population structure as follows. There are two types of agents in the model, queens and workers. Reproductive females or queens age at a given rate and have specified birth and death-rates. Offspring stay within the parent colony at a given rate. Migrating progeny can start a new colony or join another colony, and there is a specific rate at which the non-migrant offspring are successfully established as queens. Workers, on the other hand, have an ageing process with a specific death-rate. The agents interact according to the rules outlined below.

Solitary species consists only of queens which reproduce or die at a defined age-dependent rate. Monogynous and polygynous eusocial strategies result in populations that are organised in colonies with both queen and worker agents. All queens and workers are part of any one of the colonies. Each offspring of a queen can either remain in order to become workers or leave to start a new colony in monogynous populations or in polygynous populations, can also join any other colony as additional queens at a specified successful establishment rate. The reproductive rate of a queen is colony size, namely on the total number of workers and age dependent; in polygynous colonies, the reproductive rate is scaled in order to account for competition between the queens. As in previously defined models, both short and long-term strategies are employed. In the short strategy, the effective birth-rate b starts from a high value, linearly reducing to a minimum at age r, while for the long term strategy, b starts from a low value and increases linearly

(Fig 2.3a) till age r and for $a > r$ $b = 0$. To make the ageing strategies more generic, a tent-shaped function is considered. Here the h_1 and h_2 represents the birth-rate at initial and at age group r and p represents the age group at which the birth-rate is maximum. As the life time reproductive capacity (K) and number of reproductive age groups (r) is constant across finite age models, and given h_1, h_2 are also equal and constant, the value of the birth-rate $b(r)$ at any chosen p will be the same. Here p will define the ageing strategy (Fig 2.2b): smaller p indicates the short lifespan strategy and larger p the long lifespan strategy.

2.3 Results and Analysis

The models discussed above in the previous Section have a finite age-structure, with the birth-rate and intrinsic death-rate being linear functions of age. The ageing function is the difference between the age-weighted birth- and intrinsic death-rates, and is represented as effective birth-rate function. For constant life-time reproductive capacity K, the larger the age-weighted effective birth-rate, the longer is the lifespan of the organism (Fig 2.2). These ageing strategies thus can be competed and a particular dominant strategy can emerge. Note that I define the fitness of a population as its size.

2.3.1 Solitary Populations

Eqs. (1-4) and (5-7) represent the age-structured population model for solitary organisms, with the fitness given in Eq. (4). In our simulations, I take $K = r = 20$ and vary h_2. Fig 2.4 shows the resulting fitness as a function of h_2 and the death rate, D. Since the age-weighted average effective birth-rate ($\int_0^r a \times b(a)da = \frac{1}{6}r(rh_2 + 2K)$) is proportional to h_2, a larger value of h_2 would correspond to the "long" ageing strategy. As expected, independent of the value of the extrinsic death-rate D, the fitness decreases monotonically with h_2 showing that independent of the extrinsic death-rate, a short (semelparous) life history is inevitable for a simple population with the specified ageing function.

A solitary population cannot reduce the ageing rate (namely increase the lifespan) due to reduced extrinsic death-rate under the conditions of these simulations. For solitary species, the short-term strategy always outcompetes a long term strategy, suggesting that long lifespans cannot be evolved in solitary species. I have also simulated Eq. (2.6): the asymptotic population size shows (Fig 2.5) that the short strategy *always* out-competes others. As above, I use a linear variation, with h_2 defined in a similar manner: Similar results (not shown here) were achieved using agent-based model of solitary populations. This show the concurrence between our models.

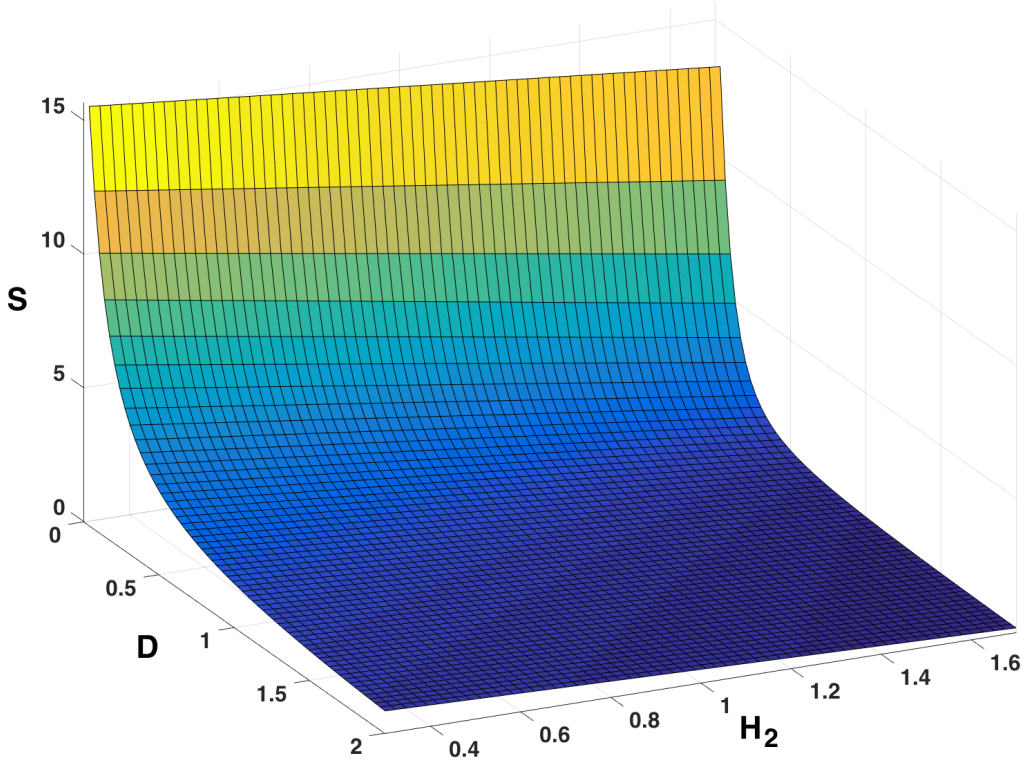

Figure 2.4: Fitness of the age-structured solitary population, plotted as a function of extrinsic death-rate D and h_2 which decides the ageing strategy (see text).

2.3.2 Eusocial Populations

Eq. (2.7-2.9) specifies the age-structured population model for monogynous eusocial organisms. As in the case of the solitary population, fitness is measured as the size of the population, namely Eq. (2.9) in steady state. The ageing function defined by h_2 (Fig 2.2a) is used to normalise the birth-rate and/or death-rate, and a number of different strategies are modelled.

Fig 2.6 shows the fitness as a function of h_2 and D, and here one can see a departure from the behaviour of solitary populations (Figs. 2.4 and 2.5). For low extrinsic death-rate, the "long" strategy has a higher fitness than the short strategy, and for high extrinsic death-rate, the opposite is true. Due to the characteristic shape of the fitness landscape, there are regions of non-monotonic fitness, namely for constant extrinsic death-rate the fitness is minimal at an intermediate value of h_2, suggesting that fitness can grow both by increasing or decreasing the lifespan. It means that in a population started with a queens of specific life-span, the decrease or increase of extrinsic death-rate might not

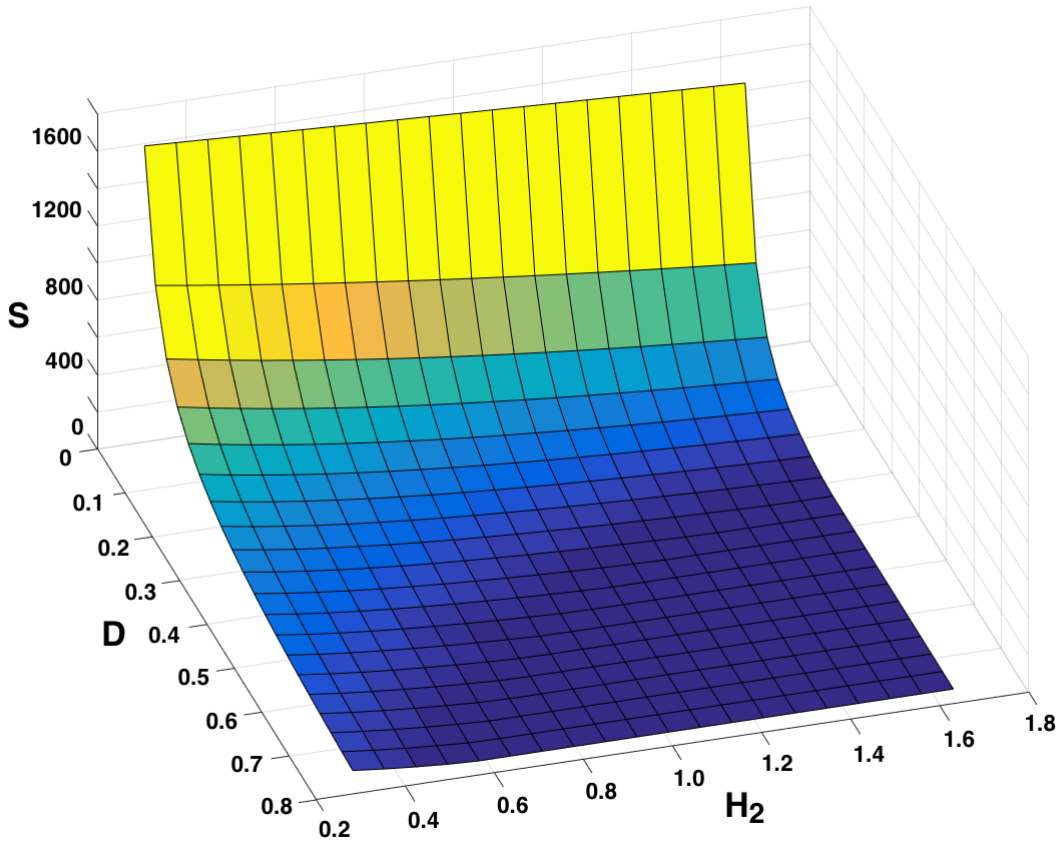

Figure 2.5: Fitness of the solitary age-structured evolutionary model, plotted as a function of the extrinsic death-rate D and h_2 for the solitary evolutionary dynamics model. As in Fig 2.4a larger h_2 corresponds to a long-life strategy.

always lead to increase or decrease in life-span respectively and contraty can be possible.

When I competed these monogynous eusocial strategies with one another (also using the agent-based model), I achieved results in accord with the above fitness landscape. When long and short monogynous strategies are competed the short strategy outcompetes the long when probability to stay q is small and the colony benefit g is large. On the other hand, the long strategy dominates for median and high values of q (data not shown). When polygynous eusocial strategies compete against each another in an agent-based model, I find that the short strategy dominates for a wider range of q than the monogynous populations. This suggests that long lifespans can also be evolved in polygynous eusocial populations although over a comparatively limited parameter range.

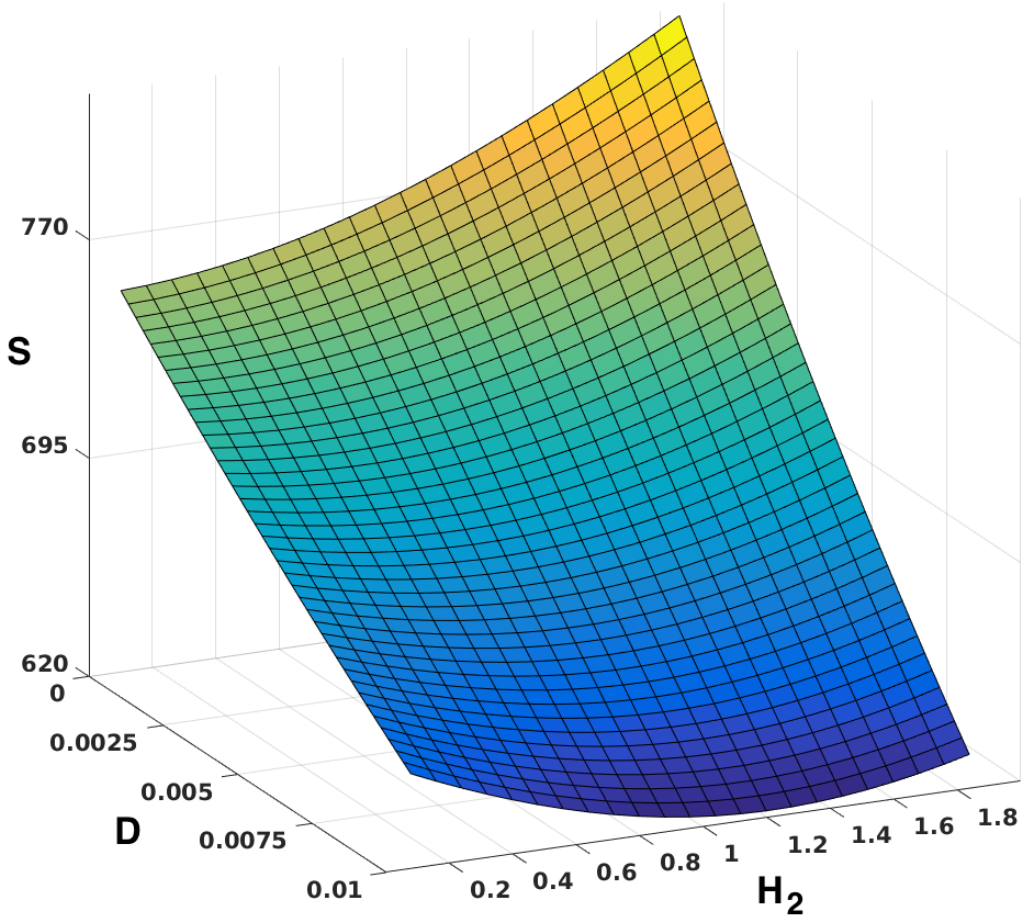

Figure 2.6: Fitness of the solitary age-structured evolutionary model, plotted as a function of extrinsic death-rate D and h_2 for monogynous eusocial evolutionary dynamics model. h_2 is proportional to the ageing strategy.

2.3.3 Solitary vs Monogynous Eusocial strategies

I competed solitary strategy with the monogynous eusocial strategy using both evolutionary dynamics model and agent-based model. The simulations are carried out with an initially equal number of solitary and eusocial individuals in the population, and the dynamics are allowed to evolve for different q, g and h_2 until one of the strategies dominate. Results are given in Fig 2.7: the dominant strategy is shown as a function of the three parameters. The monogynous eusocial strategy dominates over a larger range of q and b when long lifespans (larger h_2) are considered as opposed to the case when shorter lifespan strategies (small h_2) are considered.

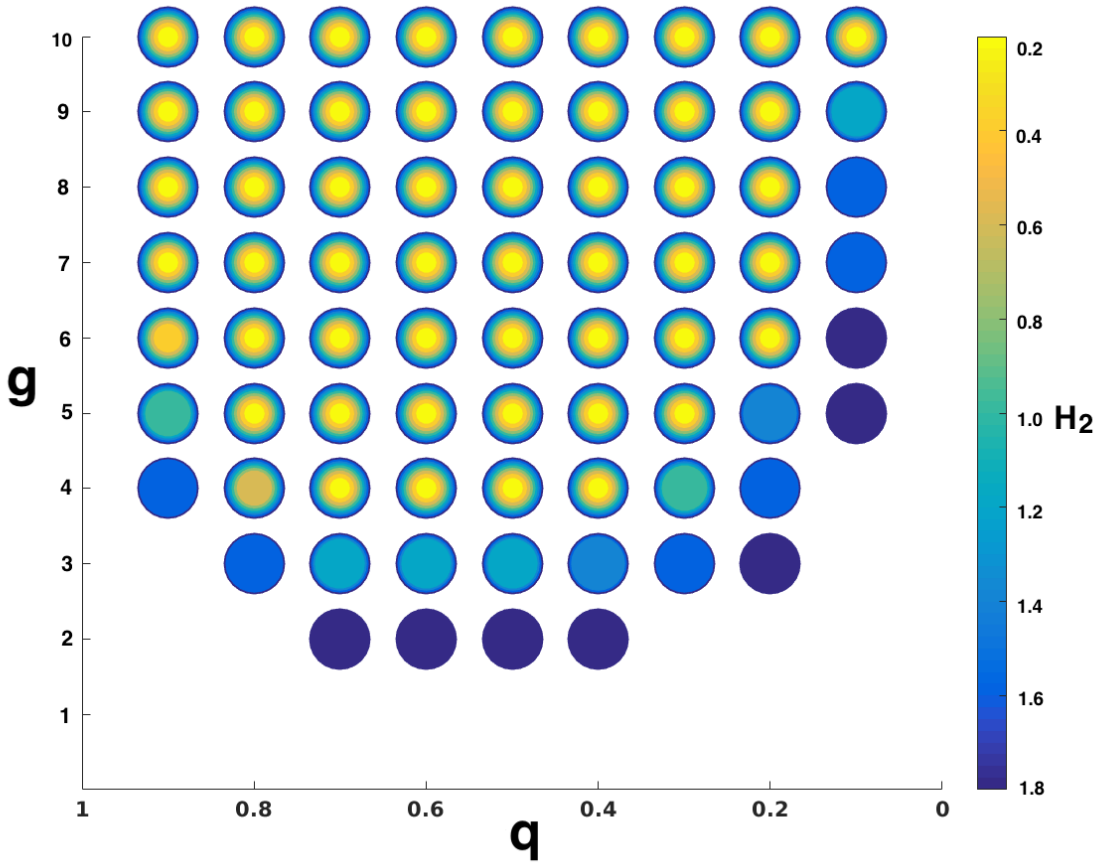

Figure 2.7: Competition between solitary and monogynous eusocial populations: Each filled circle denotes the evolution of monogynous eusociality. The colour of the filled circle denotes the value of h_2. In competition, when long strategies are employed, monogynous eusociality evolves for lower g values than when short strategies are used.

The four strategies that are possible come from combinations of solitary versus monogynous eusocial populations with monotone decreasing birthrate ($h_2 < h_1$) or increasing ($h_2 > h_1$) namely the short or long ageing strategies (Fig 2.2a): these are denoted SS, ES, SL, and EL. Simulations are carried out for different b and q with equal numbers of all four types of individuals in the population initially. The system is allowed to evolve until a single strategy dominates: this is shown as a function of the two parameters in Fig 2.8. The solitary short (SS) and EL (eusocial-long) are the only strategies that eventually dominate, and the other two possibilities, namely SL, the late-breeding solitary populations or ES, eusocial populations which breed early are not seen in our simulations. For median ranges of the probability to stay q, I find that eusocial populations with a long period of queen fecundity dominates, while solitary populations with early reproduction are preferred when the probability to stay q is close to 0 or 1. The parameter range cor-

responding to the evolution of eusociality increases in comparison to the age-independent model studied in NTW [49]. Compare Figs 2.7, 2.8 with corresponding results in [49]. Within this model, therefore, this indicates that a long lifespan *promotes* eusociality: a monogynous eusocial population with a long-lived queen almost always outcompetes a similar society with a short-lived queen unless the probability to stay (q) is low and group benefits g are high and a solitary one unless the probability to stay (q) is at extremes.

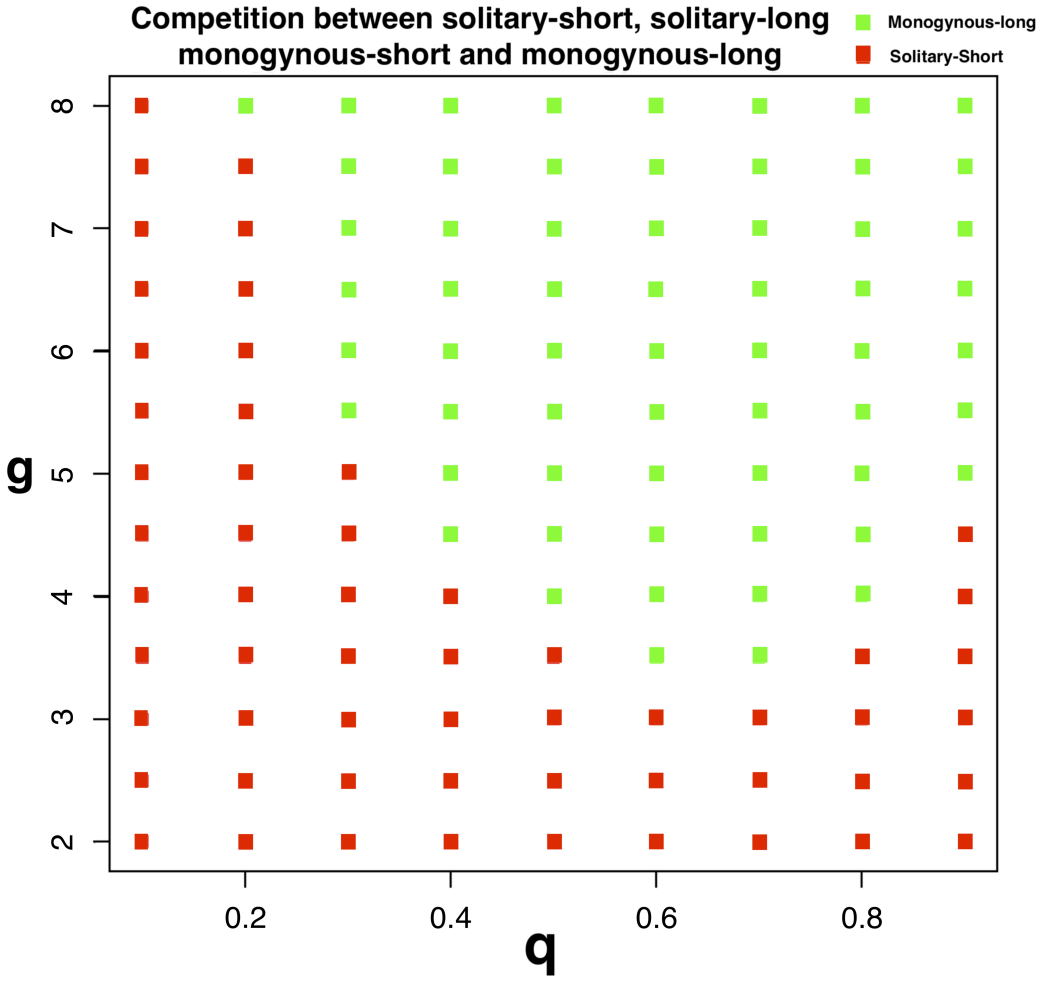

Figure 2.8: Competition between SS, SL, MS and ML strategies: Green boxes denote evolution of the monogynous long-lifespan (ML) and red boxes denote evolution of the solitary short-lifespan (SS) cases. As discussed in the text, SL and MS are not observed.

2.3.4 Solitary vs Monogynous vs Polygynous eusocial strategies

If one allows the colony to be polygynous, there are two more strategies to be considered, namely PS and PL. Using the agent-based model, all six strategies are defined and simulations are carried out as described above; For chosen b, the SS strategy dominates for both low and high q, while the polygynous strategy PS dominates for intermediate q, and the monogynous is dominant at large q values (data not shown).

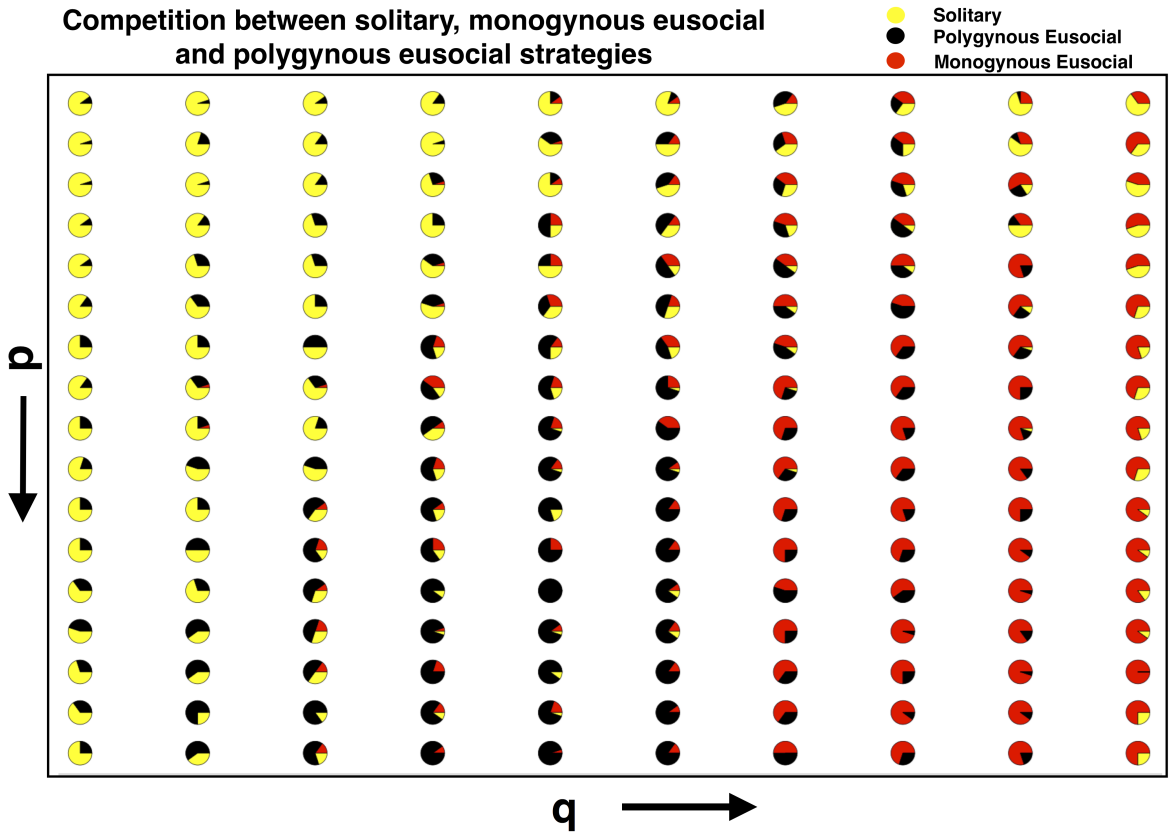

Figure 2.9: Competition between solitary, monogynous eusocial and polygynous eusocial populations: a unimodal reproduction profile is assumed, and the stay-put probability q is varied along the abscissa and the the maximum reproductive age p is along the ordinate. Monogynous eusociality, which evolves for larger values of p and q is depicted in red, the solitary strategy (in yellow) arises for lower values of p and q, while the black dots correspond to polygynous eusociality which comes about at intermediate values.

As discussed in previous section, one can consider more general age-fecundity structures (Fig 2.2b). If the rate of oviposition is taken to be maximal at reproductive age p (I took a piecewise linear function, increasing to a maximum at p with a subsequent linear decrease). Varying q and p but keeping the area under the curve constant I compete soli-

tary and both eusocial strategies in order to determine the dominant strategy. As can be seen in Fig 2.9, the solitary strategy dominates for low values of q (decreasing somewhat with increasing p). For intermediate q the solitary strategy emerges only at low values of p while the polygynous strategy dominates for larger p. For higher probability to stay, as may be expected, the monogynous strategy dominates, except for the possibility for low p when the solitary strategy may be preferred.

This would suggest that within this model, the evolution of eusociality and long lifespans, namely larger p, are correlated. Short lifespans (low p) favour the solitary lifestyle, while for long p eusociality (whether polygynous or monogynous) is preferred. For values of q where all three strategies are possible, increasing p favours the monogynous strategy. This is in accord with the observation that polygynous species can have shorter lifespans compared to monogynous species with similar extrinsic death rates.

2.4 Summary and Discussion

Classical theories of the evolution of ageing have limitations[123, 124, 125, 139, 140, 141, 142]; empirical observations show that in some cases there are departures from classical predictions [123, 124], and mechanistic details of the evolution of ageing in many populations is generally unknown [125]. Furthermore, ageing is a process of considerable complexity [143, 144, 145, 146, 147]. Several earlier studies have quantitatively shown that the evolution of eusociality and that of long lifespans are highly correlated [104] although the mechanistic details need further exploration. Newer approaches such as the study of hierarchical trade-off models [127] and the inclusion of intergenerational transfer [148] are being employed to better understand such details. I therefore decided to explore the effect of a population structure on the evolution of eusociality.

In the present work, I have built upon an evolutionary model of eusociality that implements population structure [49] to include age structure. Using an age-structured NTW model and a related agent-based model, I show that the eusocial population structure can increase the fitness of long lifespan strategies. With fixed reproductive capacity in a lifetime and with extrinsic death in addition, strategies that favour slow senescence in solitary species are evolutionarily expensive. When there are considerable group benefits that accrue for longer lifespans, eusocial species outcompete solitary species. On the other hand, when the lifespan is short solitary species outcompete eusocial ones. The analysis of the fitness landscape for eusocial populations shows that with increased extrinsic death-rate, a population can increase fitness both by increasing as well as by reducing the intrinsic lifespan.

The comparison between fitness landscapes further shows that the evolution of long

lifespans is possible over a larger region of parameter space for monogynous rather than polygynous eusocial populations. In polygynous eusocial species — intermediate between single queen colonies and solitary species — additional queens that join a group by migration effectively pull down the average age of colonies. The progeny can gain group benefits by joining another group which has already crossed a threshold size and so can increase early reproduction rate. In such cases, comparatively faster ageing (or shorter lifespans) would be beneficial. The ageing strategy also controls the number of colonies that exceed a threshold size for group benefits. Since the correlation is inverse, if the ageing rate of polygynous species is increased (namely a reduced lifespan) there can be a situation when there are no groups available which are above the threshold size. Progeny will then not be able to draw upon the benefits of joining a mature colony, and a solitary strategy would be preferred for very short lifespans. A similar argument can be made to show that for long lifespans the polygynous populations are dominant when the probability of progeny migration is large, and monogynous eusocial populations are dominant when this probability is small.

The competition simulations between long lifespan strategies show that the eusocial populations outcompete solitary population for a larger parameter space of probability to stay q and group benefits g. In our agent-based modelling simulations, the inclusion of polygyny along with the solitary and eusocial strategies allows for a better exploration of the parameter space for evolution of eusociality. Polygynous eusociality occupies an intermediary region in the phase space, between the solitary and monogynous eusocial cases. The evolution of long lifespans is thus intrinsic to eusociality, both evolving together. Indeed, it would appear from the present study that in this model a long lifespan in social organisms is both a product as well as an enabler of eusociality. This would suggest that including population structure in ageing models is important and further stresses the importance of heterogeneous modeling approaches.

In future work I intend to study extensions of the present models that can include more realistic forms of various age- and size-dependent parameters such as group benefits and death rates. A number of other features such as food availability, maturation time, or foraging time can be included, and their effects need to be explored since some of these factors are empirically known to affect ageing. Similarly, specific population and life-history structures such as the mating behaviour, maturation stage, inter-generational co-operation, metabolism and maintenance, need to be included and explored. the models themselves can be made more sophisticated by including processes such as mutation-selection.

Due to generality of the present models, specific quantitative predictions are difficult to make. However, the qualitative and theoretical results obtained here should ideally

find validation through empirical observations. Nevertheless, eusociality is always correlated with long lifespans [104], and polygynous eusocial species have a lower lifespan compared to a monogynous one for a similar extrinsic death rate [114]. Understanding the mechanisms of ageing will have profound implications in medical interventions for senescence-related damage. In addition, insights into the ageing process will have an impact on our understanding of the dynamics of populations and their evolutionary trajectories.

CHAPTER 3

"By-product" group benefits of non-kin resource-sharing in vampire bats

I develop an agent based model (ABM) to simulate the behaviour of a colony of vampire bats (Order: *Chiroptera*) and study the by-product group benefits that result from resource-sharing among related as well as unrelated members of the colony. Such cooperative behaviour can can lead to unexpected group benefits; there is an increase the inclusive fitness of related members of the colony (namely *kin*) and can have direct benefit when shared with unrelated members (namely *non-kin*). Sharing can also provides by-product benefits when individuals have a shared (or *group*) interest.

Our study focuses on the contrast in the group estimates between sharing and non-sharing populations. For constant ecological resources, sharing behaviour can increase the sustainable population size, increase the total resource stored in the population, and reduce the average resource required per individual, compared to a non-sharing population. (The extent of the increase or decrease will depend on the parameters of the model). This increased carrying capacity due to resource sharing can increase the fitness of individuals in the group. The increase in cooperativity has a nonlinear effect on group benefits: Substantial group benefits are shown only after a cooperativity threshold, and it increases exponentially to a maximum thereafter.

3.1 Introduction

Agent based modeling (ABM) techniques are known to provide considerable insight into a number of different problems in different areas of enquiry, ranging from biology, physics, and chemistry, to economics and the social sciences [149, 150, 151]. In order to understand the emergence of properties in a complex system, its parts are modelled as interacting agents with a specified minimal set of properties and behaviours. As an example of such an approach, the properties of an ecology can be seen to derive from the populations of its constituent species [152, 153] and agent-based modelling has been usefully applied in to understand the complex process of the emergence of cooperation [154] among the species. In general, ABMs using simple local interactions can give insight into complex global patterns [155]. Our aim in this work is to use autonomous behaviours of vampire bats (such as foraging, starvation, death, breeding and blood sharing) in an ABM framework in order to understand the effect at a population level, particularly in respect of the inclusive fitness of the population.

Cooperating organisms often invest in partners preferentially so as to increase the inclusive fitness benefit [156]. Inclusive fitness is the sum of direct fitness and indirect fitness [157, 158, 159, 160]. The ability to identify close relatives, namely the phenomenon of kin discrimination, and reciprocity (a tit for tat strategy) are mechanisms that ensures that so–called cheaters do not benefit from cooperation. On the other hand, when the act of cooperation automatically provides so-called by-product benefits (which in general are shared group benefits) no specific enforcement may be required [157, 161]. In this chapter I explore such group benefits.

In the common vampire bat (*Desmodus rotundus*) food or blood sharing is a cooperative behaviour [162]. Vampire bats are obligate blood feeders and can store very limited resources for survival; a 72 hour starvation will kill the bat [163]. At the same time, bats regurgitate in order to share blood with kin, namely related members of the colony, as well as with others [164]. Cheaters, namely those who do not reciprocate help can be detected by social grooming [156, 165]. It has been observed that food-sharing with unrelated members of the colony, namely non-kin, occurs preferentially with individuals having high past reciprocation [106, 162].

Simulations by Wilkinson [164] have shown that the direct fitness benefit is low compared to indirect fitness benefit. A bat with 90% success rate of foraging takes 1110 days on average to miss three consecutive meals. Thus a typical bat may need no more than

3 to 5 donations of food through sharing in its entire lifespan. Considerable attention has been given to examine both primary (direct) fitness and indirect fitness benefits and mechanisms to maintain cooperation [106, 156, 162, 164]. By-product group benefits of food sharing in vampire bats have also been studied, although to a lesser extent. This is useful in understanding cooperation among non-kin individuals, and as has been seen in simulations, energy sharing as in vampire bats can bring substantial benefits to the group as a whole [166].

I present an agent-based model (ABM) of food sharing in vampire bats. Our simulations explore the group benefits of resource sharing with all individuals in the group. I find that within our model, for a given constant rate of ecological resources, both the carrying capacity as well as the total resources stored with the individuals in the population increase significantly. This increase in the sustainable population size with a small increase in resources gathered can reduce the resources required per individual.The increase in fitness due to the increased group size can result in the increased reproductive capacity of individuals.

I have also examined the effect of the rate of cooperativity on group size and find that considerable group benefits can occur only for large cooperativity. This suggests that such by-product benefits might be of use in maintaining the cooperativity, although this cannot explain the origin of cooperativity itself. Another factor in the model is the capacity to store food by an individual: this is varied from sufficient food for 3 days to sufficient food for 12 days. The increase in capacity to store food reduces the sustainable size for both cooperating and non-cooperating populations, but the population size ratio (sharing to non sharing) is nonmonotonic, increasing first and then decreasing. Our present results suggest that significant additional by-product group benefits accrue from food-sharing behaviour in vampire bats, and this aspect needs to be explicitly included in any estimation of the total fitness benefit of such cooperative behaviour.

3.2 Materials and Methods

The agent based model that I employ here is as follows. Each individual bat forages every day, and if successful will store three units of food resource. The success is probabilistic and depends on its own efficiency and resources available per individual. The rate of food resource in the habitat (total resource per day) is taken to be constant and thus the resources available per bat per day is inversely related to the population size. If foraging is unsuccessful the bat's stored resource is reduced by one unit, and the particular bat

survives only if the stored resources have a non-zero value. This is in keeping with the observation that a bat that is unsuccessful in foraging for three consecutive days is not likely to survive [162, 163]. The foraging efficiencies for initial individuals is randomly chosen in a range (for convenience between 1/2 and 1) and I further assume that each bat reproduces once every year at a random time [167], offspring inheriting the mother's foraging efficiency.

In a population with resources-sharing behaviour, individuals that are successful in foraging are considered donors and are denoted D: these have at least 3 units of stored food. Bats with only one unit of stored resources, namely those that are unsuccessful in foraging, are denoted N (for needy). So long as the total number of N bats is lower than the D, each will receive one unit of resource randomly from a donor. If N exceeds D, then each donor randomly selects one N bat to donate one unit of resources.

I find that in our simulations, the model bat populations stabilize after a transient time (which depends on the efficiency range chosen). Quantities such as the stable size of the population, the total food stored with individuals in each population and the resources available per individual can be measured, and compared between populations that indulge in food-sharing versus those that do not. For proper comparison, I take both populations to feed on the same amount of external resource, to have the same foraging efficiencies, and further, that both groups start with the same number of individuals. The efficiency of a strategy is measured by the size of the stable population that is eventually achieved.

In order to understand the effect of partial cooperativity, each resource-sharing event is taken to be probabilistic. If the index of cooperativity is w, only that fraction of resource-sharing acts will be successful, while others maintain the *status quo* by not sharing. Another parameter of interest is c, the capacity of an individual to store resources, and in our simulations I have varied its value from three to twelve. The model, data and NETLOGO code for reproducing the results can be downloaded from the https://github.com/Donepudiraviteja/Resource-sharing/.

3.3 Results

I present results in Fig. 3.1 for populations of equal size, with each habitat having the same rate of resource availability. Transients are discarded and the different measures are calculated after the populations have stabilized.

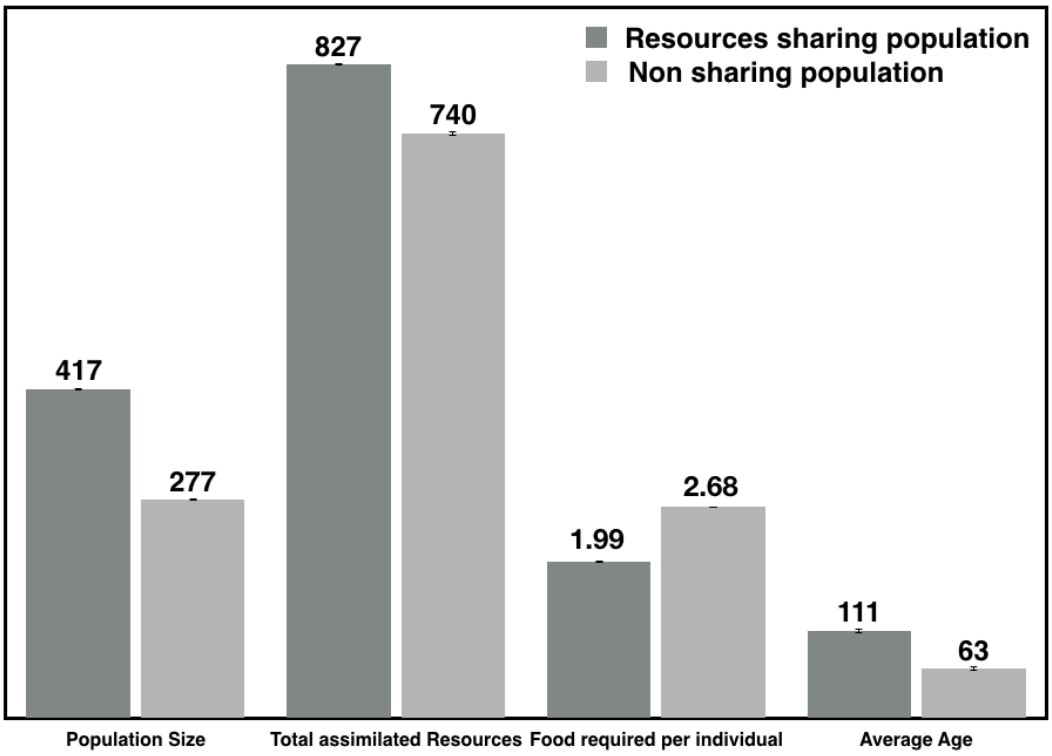

Figure 3.1: Comparison of measured quantities for the sharing and non-sharing populations, starting with 200 individuals and 600 units of ecological resources. A transient time of 100 days is taken, and simulations have been averaged and rounded over an ensemble of 12 realizations. The sample to sample variation is small and does not show up on this scale.

The sustainable surviving population on the given constant food resource in the habitat for population with resources sharing behaviour is 417, where as for non-sharing behaviour it is 277 for this set of parameters, indicating that on average, the sustainable population size increases by approximately 50% due to resource-sharing, since this is the only difference between the two models. These results are typical: simulations for various levels of resource availability per day showed similar trends: both sharing and non-sharing populations have a linear population growth as a function of the resource availability, and with a similar size ratio (data not shown).

The total resources assimilated from the habitat in a single day is calculated by adding the resources available to all individuals. In our simulations, sharing behaviour actually increased the total resources assimilated by about 11% over non-sharing population since the cooperative behaviour allowed for more foragers and consequently, fewer N individ-

uals. The resource required per individual, namely the total food resources assimilated divided by the population size is about one fourth *less* for food sharing population.

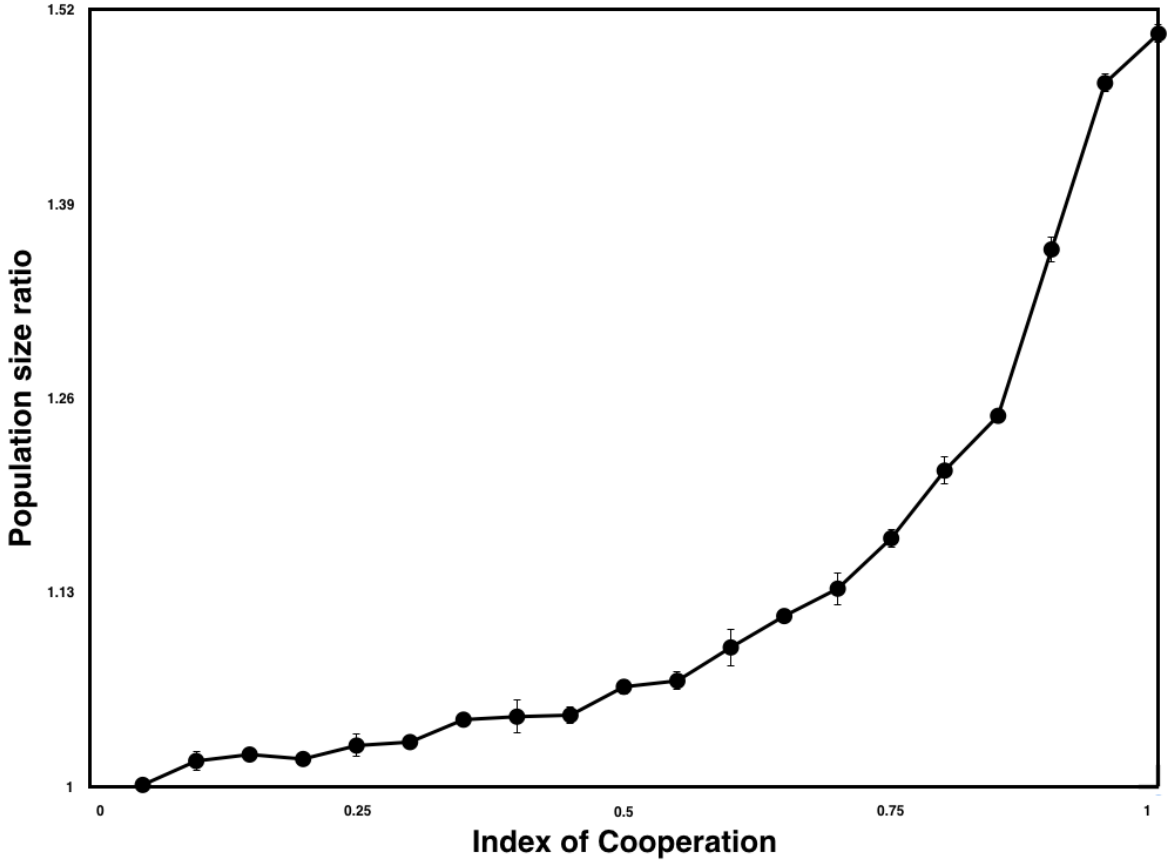

Figure 3.2: Group benefits with increasing cooperativity. The cooperativity of the population is varied on the abscissa, and the resulting population size ratio (between populations of sharers versus non-sharers) is on the ordinate. All results are averaged over 5 realizations, and the error bars are shown.

I also considered the case when individuals were pre-classified as resource sharing or non-sharing, and allowed them to compete for resources. An equal number of both types of individuals was considered, and in steady state, I find (see Fig. 3.2) that resources sharing individuals eventually take over the population: this is the dominant strategy. I keep the populations distinct, namely there are no resource transfers between sharing and non sharing individuals, and there are no defectors between the types in the population. Simulations show that an increase in cooperativity leads to a nonlinear increases in group benefits: substantial benefits are shown only after the cooperativity parameter is 60%, and it increases exponentially to a maximum thereafter.

The sustainable group size is strongly dependent on the maximum capacity c that an

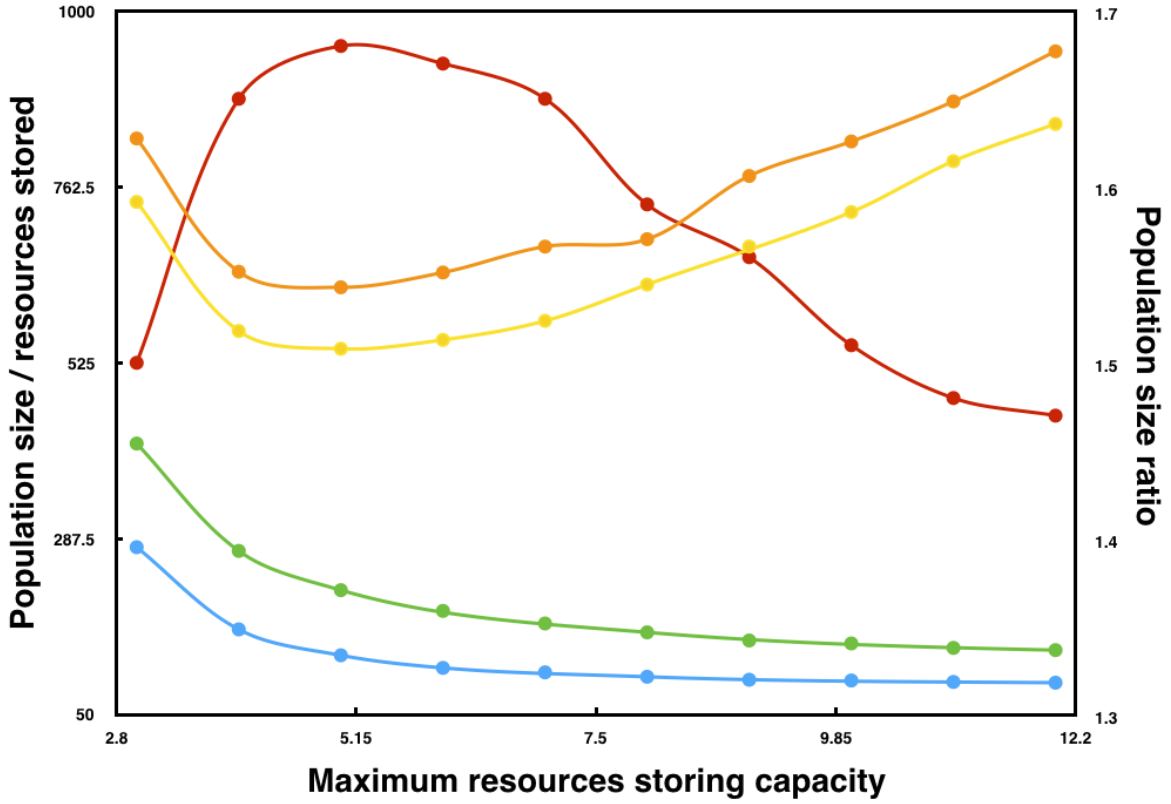

Figure 3.3: Group benefits with increasing maximum resource storing capacity, c plotted along the abscissa. Red curve represents population size ratio between sharing and non-sharing populations, blue and yellow curves represent the population size and total resources stored for non sharing population. The green and orange curves represents the population size and total resources stored for a sharing population. All results are averaged over 5 realizations.

individual possesses to store resources. Group sizes were measured for c varying between 3 and 12. The effect of increasing capacity is, paradoxically, to *reduce* group size in both sharing and non-sharing populations (see Fig. 3.3). The ratio between sharing and non sharing populations has a nonmonotonic dependence on c whereas the total resources stored by both sharing and non sharing populations has the inverse effect.

Finally, to see whether the increased population size effects the lifespan, I measure the average age and the average age at death for both the populations. (In this simplified model, I only consider death to occur via resource-deprivation). The food-sharing population, on average, have age 1.8 times as the non-sharing population and the average age at death for food-sharing population is twice that of non-sharing population. If p is the probability of unsuccessful forages the average number of days before there are three

consecutive unsuccessful forages (TCUF), is clearly given by

$$\text{TCUF} = \frac{1 + p + p^2}{p^3}.$$

For the reported probability of 10% unsuccessful forages [162, 164], therefore, on average TCUF will occur after more than three years; the average lifespan of a vampire bat being about 10 years.

3.4 Summary and Discussion

Pseudo-reciprocity [161, 168] results when the cheaters (in the sense used here) indirectly benefit those individuals who share resources. This is a *by-product* group benefit, and in the present work I study a model of resource-sharing between individuals in a bat colony where such behaviour can lead to substantial group benefits. Mechanisms such as kin recognition and reciprocality ensure and reinforce cooperative behaviour. The by-product group benefits that result in our simulations has the effect of increasing the effective total fitness of individuals and thereby might also help in maintaining the fitness through pseudo-reciprocity.

Increasing the maximum capacity of resources that can be stored affects the quantifiers I have focussed upon for both sharing and non-sharing populations. However, the ratio of their respective population sizes is a nonmontonic function of storage capacity, suggesting that for a species that can store resources for sufficiently long, there is unlikely to be any added benefits due to cooperation.

Cooperativity has a nonlinear effect on benefits. For low levels of cooperativity the benefits are also low, but after the cooperativity increases to about 60% benefits rapidly increase. In the evolution of food-sharing behaviour, in the initial stages, when the co-operativity is low, the by-product benefits would be minimal, but once significant co-operativity is established, the consequent benefits help in its maintenance.

In summary, our simple application of agent based models to understand cooperative behaviour of vampire bats suggests that resource-sharing between unrelated individuals in vampire bat colonies can yield substantial by-product benefits. Such behaviour has not been explored in previous studies, and this suggests that more specifically defined resource-sharing models need to be developed and studied, particularly to understand better the differential advantages of various strategies of foraging and resource allocation.

CHAPTER 4

The collective dynamics of NF$-\kappa$B in cellular ensembles

The transcription factor NF$-\kappa$B is a crucial component in inflammatory signalling. Its dynamics is known to be oscillatory and has been extensively studied. Using a recently developed model of NF$-\kappa$B regulation, I examine the collective dynamics of a network of NF$-\kappa$B oscillators that are coupled exogenously by a common drive (in this case a periodically varying cytokine signal corresponding to the TNF molecule concentration). There is multistability owing to the overlapping of Arnol'd tongues in each of the oscillators, and thus the collective dynamics exhibit a variety of complex dynamical states. I also study the case of a globally (mean field) coupled network and observe that the ensemble can display global synchronisation, cluster synchronisation and splay states. In addition, there can be dynamical chimeras, namely coexisting synchronised and desynchronized clusters. The basins of attraction of these different collective states are studied and the parametric dependence in the basin uncertainty is examined.

4.1 Introduction

Rhythmic phenomena in biological systems occur on a wide range of time-scales, from the millisecond level firing in neuronal systems to cycles of decades or more that are seen in ecology [169, 170, 171, 172, 173, 174]. These rhythms are maintained by biological oscillators of various kinds [175], and their mutual interactions have been of considerable recent interest. Numerous phenomena at the cellular and subcellular levels are oscillatory, and many forms of control and signalling within biological cells are known to depend on periodicities in the expression levels of key molecular components [173, 174, 175, 176, 177].

Our interest in this work is on the *collective* dynamics of a group of oscillators that model the intracellular dynamics of an ubiquitous transcription factor, NF$-\kappa$B [178, 179]. Since stochasticity and nonlinearity are both very significant in a cellular environment, I examine a model of NF$-\kappa$B dynamics that is known to have complex chaotic dynamics and investigate emergent behaviour in an ensemble of such systems.

The transcription factor NF$-\kappa$B has a profound impact at the cellular and tissue levels, and has a significant effect at the physiological level as well [178]. Diverse signals appear to be mediated via NF$-\kappa$B: cell survival, cell proliferation [180], development [181], the maintenance of the immune system [182], inflammation response [183], and so on. These phenomena clearly involve the concerted effects of both intracellular oscillators and intercellular communication.

These aspects motivate the present study of dynamics in an ensemble of coupled NF$-\kappa$B networks. Most natural systems are not isolated, and thus one motivation for the present work is to examine the effect of coupling on the dynamics. Indeed the case of coupled networks (or even a network of networks) is likely to be the natural setting, when the interaction between systems leads to their having a high degree of dynamical correlation.

At the same time, studies on the dynamics of nonlinear systems have shown that there can be several forms of complex individual as well as collective dynamics. For instance, there may be more than one attractor of the dynamics so that different asymptotic states can coexist. Bi- (or multi-) stability can arise autonomously [184], or can result from the coupling [34, 185]. When the multistability is induced, the coupled system can display a variety of novel dynamical behaviours, ranging from global synchrony to cluster synchrony and splay states [39], some of which have found parallels in biological dynamics [38, 186]. Indeed, in such cases synchrony may well be essential for their proper functioning [187].

This chapter is organized as follows. The model NF$-\kappa$B network studied here is described in detail in Section 2. Developed over the past few years by Jensen, Krishna and co-workers [188, 189], this model has been studied in great detail; see [190] for a recent summary. All essential molecules that are involved in the network have been incorporated in this somewhat minimalistic model (details of which are given below). I explore the dynamics of a single regulatory network and describe the occurrence of multistability, the basins of attraction of the different attractors and their geometry. The collective dynamics of an ensemble of such externally coupled networks is discussed in Section 3, followed by a study of globally mean-field coupled NF$-\kappa$B networks. I conclude in Section 4 with a discussion and summary of our main results.

4.2 NF−κB model dynamics

Jensen and Krishna [189, 190] introduced the following reduced model of the NF−κB network that captures the essential features of the dynamics. It consists of a set of five coupled differential equations,

$$\begin{aligned}
\dot{x} &= V_x(N_x - x)\frac{K_z}{K_z + z} - V_z z\frac{x}{K_x + x} \\
\dot{y} &= \Gamma_y x^2 - \Delta_y y \\
\dot{z} &= \Gamma_z y - \Delta_z u(N_x - x)\frac{z}{K_z + z} \\
\dot{u} &= \Gamma_u \tau w - \Delta_u u \\
\dot{v} &= \Delta_u u - V_v v\frac{K_A}{K_A + A_{20}\tau} \\
\tau &= A_0 + A\sin\left(\frac{2\pi}{T}t\right) \\
w &= N_{uv} - u - v
\end{aligned} \tag{4.1}$$

wherein the variables represent concentrations of select components of the NF−κB module. In addition to the NF−κB protein, these include the inhibitor IκB, the tumor necrosis factor (TNF) and the enzyme IKK, which is a kinase; see Fig. 1 for a schematic of the network.

I briefly discuss the role of the key molecular species that are involved in the network. TNF is a cytokine with both pathological and physiological functions [191, 192, 193]. While its primary role is to stimulate inflammatory response during infection, but it also plays role in apoptosis as well as necrosis [192, 194]. TNF binds to the receptors TNFR1 and TNFR2 which are produced in many tissues. This binding leads to conformation changes that activate the NF−κB and MAPK pathways [192] and signalling of apoptosis. The cellular protein IκBα binds to NF−κB dimers to sterically prevent NF−κB from entering the nucleus [195]. In order that the NF−κB protein moves into the nucleus where it can carry out its function, the sequential phosphorylation, ubiquitination, and degradation of IκBα is essential [195, 196]. Some of these latter actions such as the phosphorylation of IκB and NF−κB proteins are accomplished via signals from NF−κB activating stimuli such as the IκB kinase (IKK) that is composed of the two serine-threonine kinases (IKKα and IKKβ) and a regulatory subunit (IKKγ) [196]. Clearly all these molecules play a crucial role in the nuclear localisation and activation of NF−κB.

The notation is as follows [189]: x, y, and z are respectively the concentrations of the nuclear NF−κB protein, I-κB RNA and I-κB. The external TNF concentration τ is periodically modulated with time period T. This affects the dynamics of the neutral

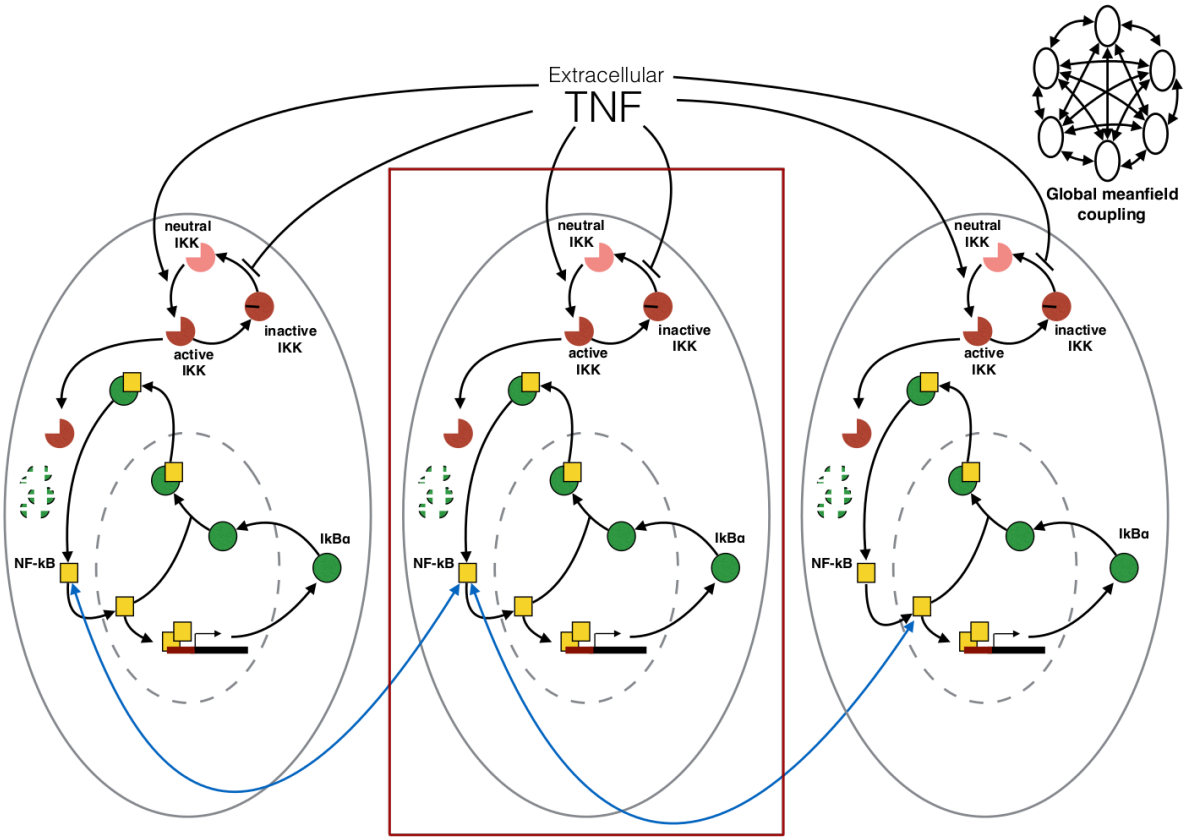

Figure 4.1: Schematic of an individual NF−κB network (see the central panel) coupled to other similar networks. Within the cell nucleus NF−κB promotes the production of IκBα which then binds to NF−κB, generating negative feedback. External TNF (common to all cells) drives IKK oscillations as indicated. In its active form, IKK in the cytoplasm helps to degrade bound IκBα, freeing up the NF−κB which is then transported into the nucleus. These coupled feedback loops effectively cause NF−κB oscillations. I also consider a global mean-field coupling through NF−κB itself (denoted by the blue arrows) which connects all the networks. Each cell is coupled to all other cells, as indicated schematically in the top right corner inset.

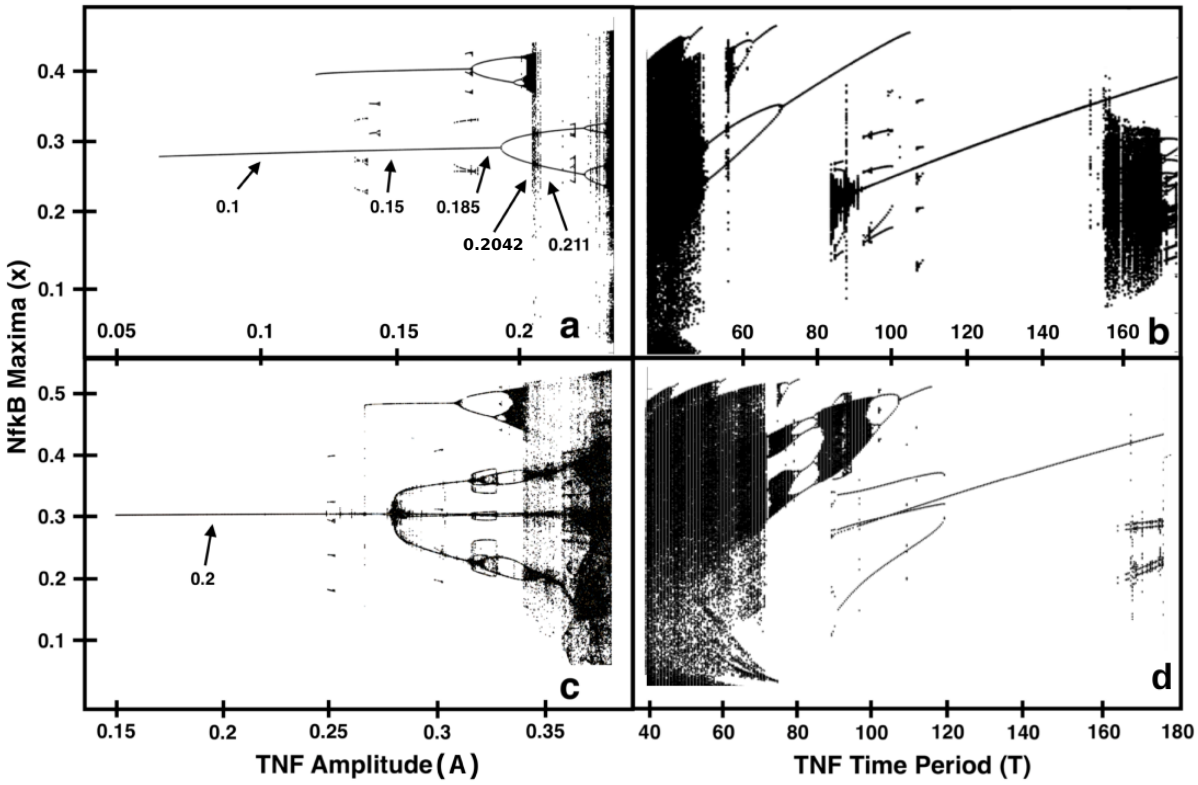

Figure 4.2: Bifurcation diagram as a function of the TNF amplitude for periods (a) T=60 (top left), and (c) T=120 (bottom left), and as a function of the period T for fixed amplitudes (b) A=0.211 (top right) and (d) A=0.3 (bottom right). Plotted on the ordinate in all panels are the maxima of the NF−κB concentration, namely x, over a sufficiently large number of oscillations.

IKK w through its action on the active and inactive forms of IKK, denoted u and v respectively. As can be noted, the above equations describe a feedback loop in (x, y, z) coupled to the (u, v, w) system that is externally driven by the autonomous and periodic TNF oscillator.

The dynamics that results from the above five equations has been extensively studied by Jensen and co-workers in a series of papers [37, 190, 197, 198]. It was recently shown [190] that depending on the modulation, there can be bistability in the system as a consequence of the overlapping of Arnold tongues: different initial conditions can correspond to different resonances.

I recall the main dynamical features of this system [37, 189, 199]. It is convenient to examine a bifurcation diagram that results from a variation of the parameters of the external modulation, namely A and T, the amplitude and period of TNF variation respectively. The bifurcation diagram is obtained in the standard way: the equations of motion are integrated, transients are discarded and a large number of successive maxima of the NF−κB oscillations, namely the variable x are plotted subsequently. Several initial

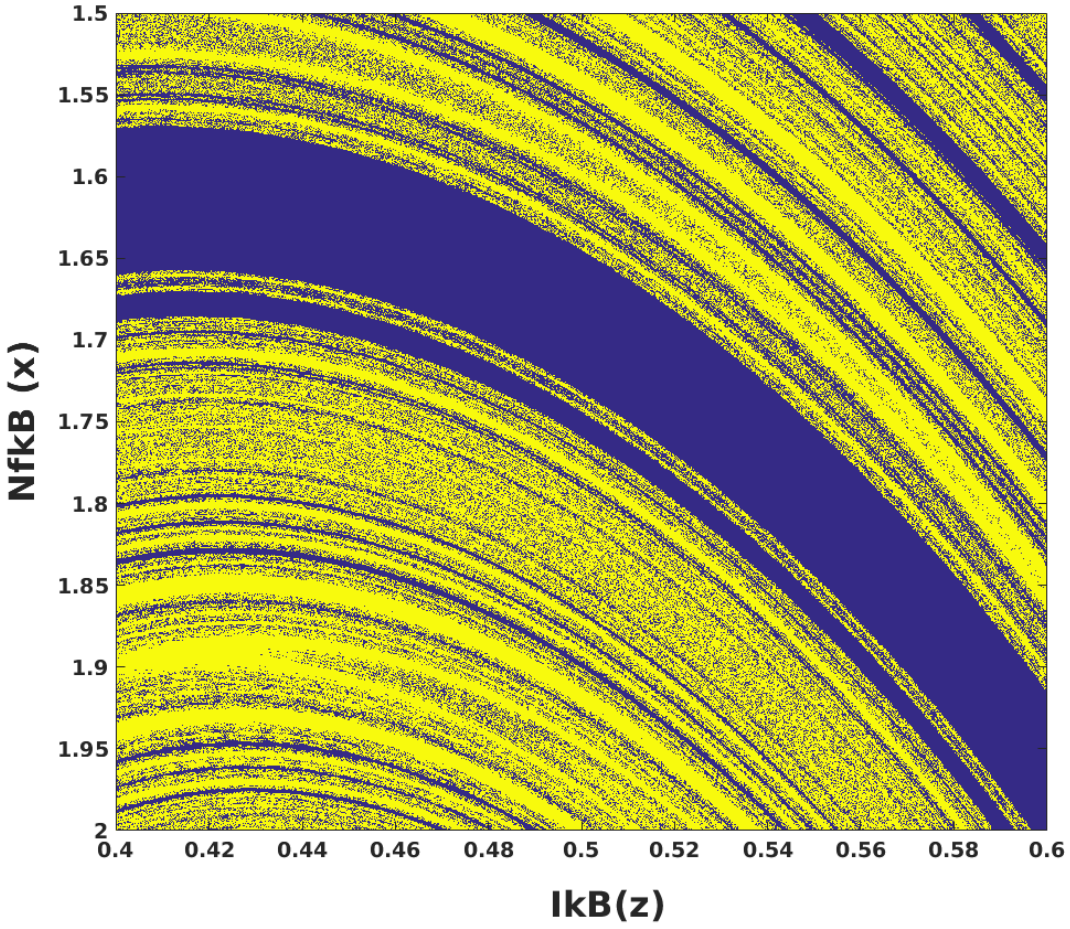

Figure 4.3: Basins of attraction for the two coexisting attractors at $A = 0.2042$ and T=60. The blue points are attracted to a chaotic attractor while the yellow points go to a limit cycle.

conditions are chosen for each value of the parameters so that bistability or multistability is immediately apparent. Fig. 4.2, which is representative, has the following features. When the period is fixed and the amplitude is varied (panels on the left), there are extensive regions of multistability that occur subsequent to period-doubling bifurcations and these appear to be of two basic types: there can be coexistence of periodic orbits (namely limit cycles) of different periods as well as periodic orbits coexisting with more complex dynamics, namely chaotic orbits. Very similar behaviour is seen if the amplitude is fixed and the period is varied (the panels on the right; the choice of A=0.211 and A=0.3 is representative). In our simulations I follow earlier studies [190] and use the following values for the several parameters in the model: $V_x = 5.4$, $N_x = 1$, $K_z = 0.035$, $V_z = 0.018$, $K_x = 0.029$, $\Gamma_y = 1.03$, $\Delta_y = 0.017$, $\Gamma_z = 0.24$, $\Delta_z = 1.05$, $\Gamma_u = 0.24$, $N_{uv} = 2.0$, $\Delta_u = 0.18$, $V_v = 0.036$, $A_{20} = 0.0026$, $K_A = 0.0018$, and A_0=0.5.

I have examined the dynamics at several values of the period T and find that there typically is a region where two limit cycles coexist. In addition, there can be exterior crises, leading to a chaotic region. The basins of attraction in the region of multistability appears to be highly mixed (I discuss the characterization below). For $A = 0.2042$ and T=60, there are two coexisting limit cycles owing to the overlapping Arnold tongues [190]. Points leading to the different attractors are shown in blue and yellow in Fig. 4.3, for a small patch of initial conditions on the $x - z$ plane.

4.2.1 Attractor basins: Boundary Fractality and Entropy

When the dynamics displays bistability or multistability, the nature of the attractor basins is naturally of interest. The manner in which the basins are arranged in the phase space has a major effect on the dynamics of an ensemble of such oscillators, especially when they may be described as being intertwined or intermingled [35, 200, 201]. In such cases the prediction of the eventual attractor for a set of initial conditions can be difficult, and a recently proposed measure, the basin entropy, attempts to quantify this uncertainty [202].

Following the procedure specified by Daza *et al.* [202], the basin entropy is computed by first classifying a set of points inside a box of initial conditions by their corresponding final states. I thus determine the number of initial conditions going to each attractor, and this is then used to calculate the Boltzman entropy

$$S_i = -\sum_{j=1}^{m_i} p_{ij} \ln p_{ij},$$

where m_i is the number of attractors in the i^{th} box and p_{ij} is the probability of the j^{th} attractor in that box. The basin entropy at a particular scale ε is the average of S_i for all boxes, namely

$$S_b(\varepsilon) = \frac{1}{N} \sum_{i=1}^{N} S_i.$$

The uncertainty coefficient α, the exponent in the scaling behaviour of S_b,

$$S_b(\varepsilon) \sim \varepsilon^{\alpha}$$

can be numerically estimated as the slope in a log-log plot of S_b versus ε. Similarly the fractal dimension d_b is determined from the scaling of the fraction F_b of points in the boxes, $F_b(\varepsilon) \sim \varepsilon^{-d_b}$.

I find that in this system, the basin uncertainty increases as a function of the TNF amplitude: see Fig. 4.4 and Fig. 4.5 (right panels). This indicates that the dynamics

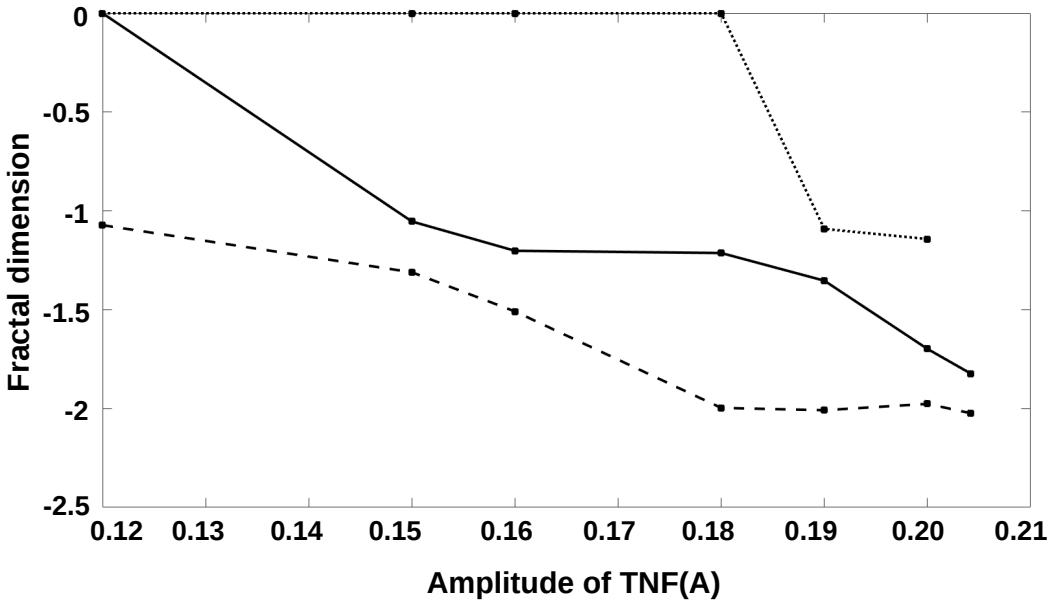

Figure 4.4: Fractal dimension of the basins as computed from the basin entropy, as a function of the TNF period and amplitude. Calculations were done for selected amplitudes and periods T= 45 (dashed) , 60 (solid) and 70 (dotted line).

of an ensemble is likely to become more complex as the external TNF amplitude is increased whereas there is a decrease in entropy when the period of modulation is increased. This correlates with earlier experimental observations by Kellog and Tay [199] that reduced doses and large periods of TNF leads to synchronous oscillations and improved entrainment. As seen in our simulations, when A is low and T is large, namely the TNF oscillations have a small amplitude and large periods, the attractor basins are well separated and would therefore be more *immune* to noise-induced hopping, leading to better entrained and synchronous populations.

4.3 Dynamics of coupled NF−κB networks

I first examine the dynamics of an ensemble of NF−κB networks which are all indirectly coupled through a common TNF oscillator: see Fig. 1. The equations of motion for this system are

$$\begin{aligned} \dot{x_i} &= V_x(N_x - x_i)\frac{K_z}{K_z + z_i} - V_z z_i \frac{x_i}{K_x + x_i} \\ \dot{y_i} &= \Gamma_y x_i^2 - \Delta_y y_i \\ \dot{z_i} &= \Gamma_z y_i - \Delta_z u(N_x - x_i)\frac{z}{K_z + z_i} \end{aligned} \tag{4.2}$$

with the subscript $i = 1, 2, \ldots, N$ labeling the individual oscillators. The dynamics of TNF is governed by the equations of motion for the variables u, v, w and τ and these are as in Eq. (1). When the parameters are such that there is a single stable attractor, all the oscillators in the ensemble synchronise into a single cluster (Fig. 4.5(a)). It can also happen that splay states are formed, namely the oscillators separate into multiple groups that are phase locked with respect to each other. Fig. 4.5(b) show two clusters denoted A_1, and A_2. The dynamics of each oscillator is on the same period one limit cycle attractor, but the two clusters are phase-locked with respect to each other. Similarly I can find splay states for period 2 limit cycle and so on (Not shown).

In addition to complete synchronization and splay states, in regions of multistability where all the attractors are not chaotic, the oscillators can also show cluster synchronization, with different subsets being attracted to different attractors. Fig. 4.5(c) shows such a state, with blocks named A_1 and A_2 corresponding to the splay states of the period one attractor while B_1, B_2, B_3, B_4, B_5 and B_6 correspond to the splay states of the period two oscillator. Similarly I can find such cluster synchronised populations for various combinations of periodic limit cycles.

In multistable regions where chaotic and periodic attractors coexist, the ensemble breaks up into clusters of in-phase or phase locked synchrony coexisting with a group

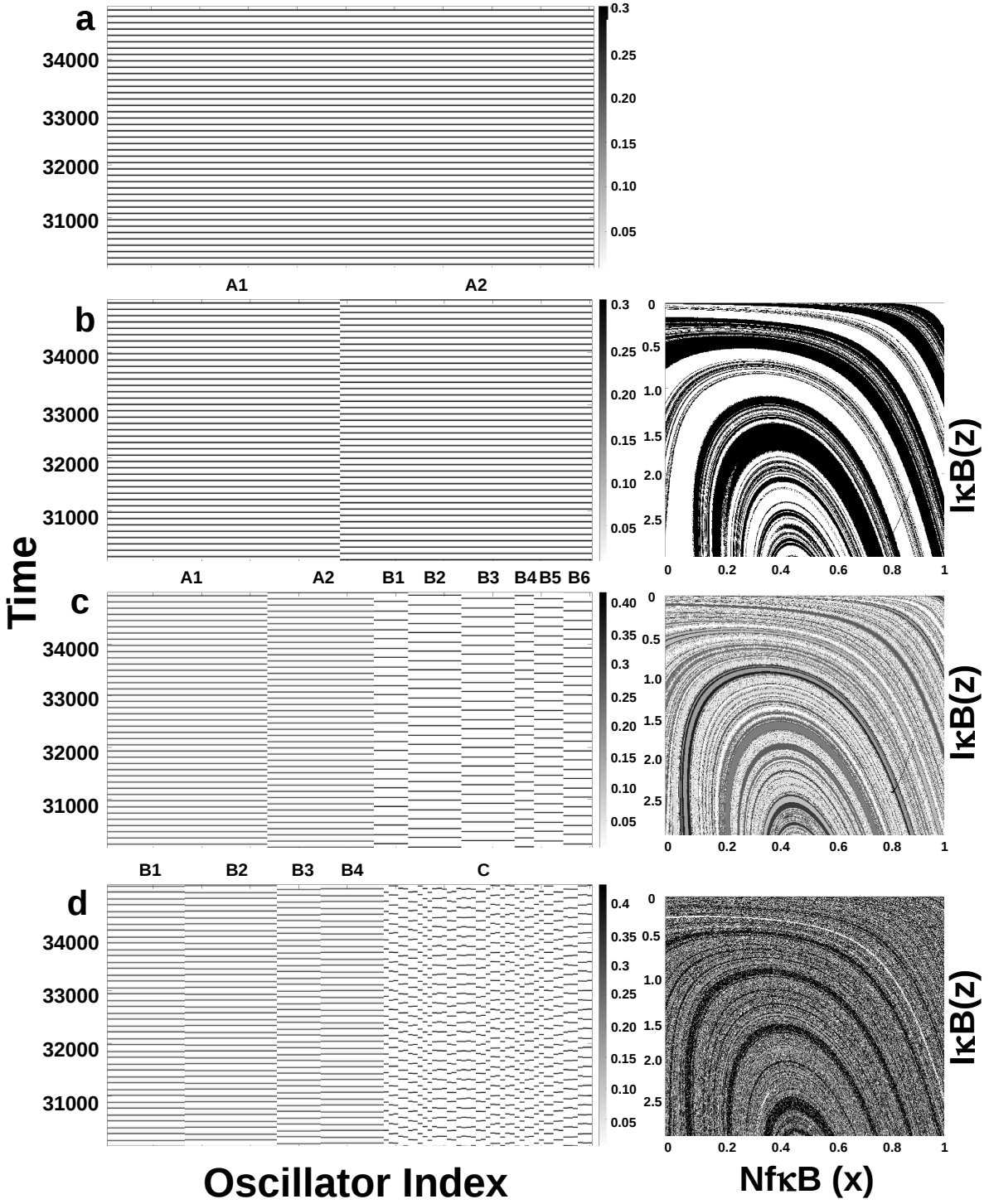

Figure 4.5: Dynamics in the coupled networks of N=100 NF−κB oscillators. The values of the TNF amplitude and period are a) T=120, A=0.2, b) T=60, A=0.15, c) T=60, A=0.185, and d) T=60, A=0.2042 respectively. The basins of attraction for the different states are shown on the right, while on the left is a space-time plot in a time interval. The maxima of the different oscillators are shown, and one can clearly identify the different clusters that are formed. Note the coexistence of coherent clusters and an asynchronous set in d).

that is desynchronised. In Fig. 4.5(d), the blocks marked B_1, B_2, B_3 and B_4 represent the synchronised groups on period 2 attractors. All oscillators in the group denoted C are desynchronised and the dynamics is chaotic. Such complexity is not exceptional in this system: chimeric states of this general type, namely mixtures of synchronised and desynchronous chaotic dynamics can be found for a range of parameters.

The emergence of chimeras in an ensemble of identical oscillators forced by a common drive was also observed in a previous study of an ensemble of Lorenz oscillators [39], where the driving itself created multistability. Here, on the other hand, multistability is a consequence of the overlapping of the Arnold tongues and this leads to chimeras in the ensemble.

4.3.1 Globally coupled NF−κB oscillators

I now consider the case when all oscillators interact through a global mean-field diffusive coupling in NF−κB . The dynamical equations are thus modified to

$$\dot{x_i} = V_x(N_x - x_i)\frac{K_z}{K_z + z_i} - V_z z_i \frac{x_i}{K_x + x_i} + \frac{K}{N-1}\sum_{j\neq i}(x_j - x_i) \tag{4.3}$$

where K is the strength of coupling between oscillators, N is total number of oscillators and the subscripts i and j label the individual oscillators.

The ensemble dynamics can be significantly affected at even low coupling as can be seen in the bifurcation diagram, Fig. 4.6. The variation with A is similar in overall structure to the uncoupled case, although new bifurcations can clearly be created. This differential influence of the coupling can lead to the establishment of newer multistable regions and also their loss. One such instance occurs at $A = 0.209$, where the coupling leads to the creation of a chaotic attractor in addition to the existing period two attractor. At this point of multistability, the ensemble shows a *weak* chimera [203], namely a partial frequency synchronisation in chaotic orbits. The oscillators on the period two limit cycle and chaotic attractor are organised into splay states. In Fig. 4.7(c) the blocks denoted by B_1, B_2, B_3 and B_4 represent splay states of the period two attractor while the dynamics in the cluster marked C_0 is desynchronised chaotic motion. A scatter plot of NF−κB maxima between all possible pairs of oscillator groups shows the difference between phase synchronisation, splay states and weak chimeras.

To understand the effect of increasing coupling I coupled a population of 100 oscillators by mean-field coupling and observed the population behaviour at various coupling strengths. I chose $A = 0.209$ where there is a single period 2 attractor and population is spliced into 4 synchronised groups, with two peaks at around 0.33 and 0.29 (Fig. 4.7(a) and 4.8(a)). I started with a very low normalised coupling constant of $K = 0.001$, the

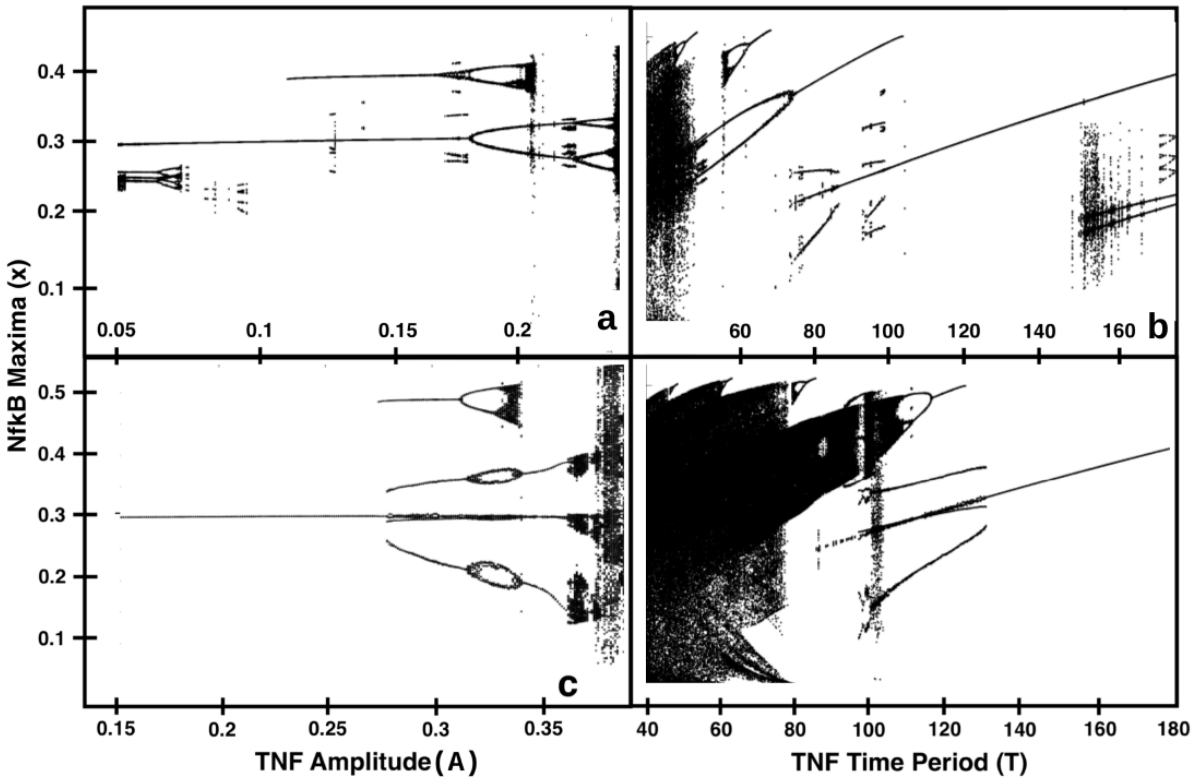

Figure 4.6: Bifurcation diagram for mean-field coupled NF−κB oscillators with K= 0.005 as a function of the TNF amplitude for fixed periods, T=60 (top-left) and T=120 (bottom-left), and as a function of the TNF Period for fixed amplitude A=0.2011 (top-right) and A=0.3 (bottom-right). The ensemble contains N=64 oscillators.

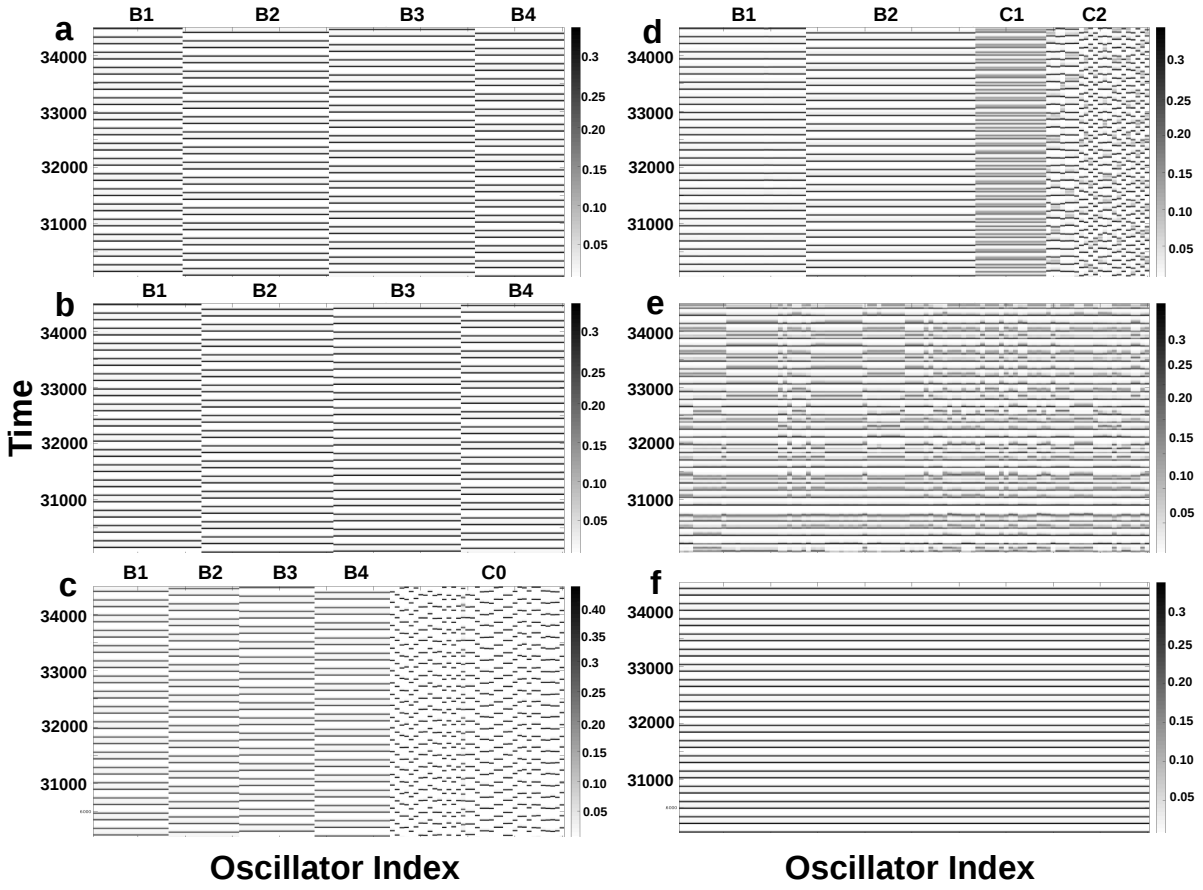

Figure 4.7: Heat plot (as in Fig. 4.5, left panels) for an ensemble of 100 globally coupled oscillators for TNF period T=60 and amplitude A=0.209. The different panels correspond to coupling a) K=0, b) K=0.001 (cluster synchrony), c) K=0.005 (chimera), d) K=0.05 (chimera), e) K=0.5 (asynchrony), and f) K= 2 (full synchrony) respectively.

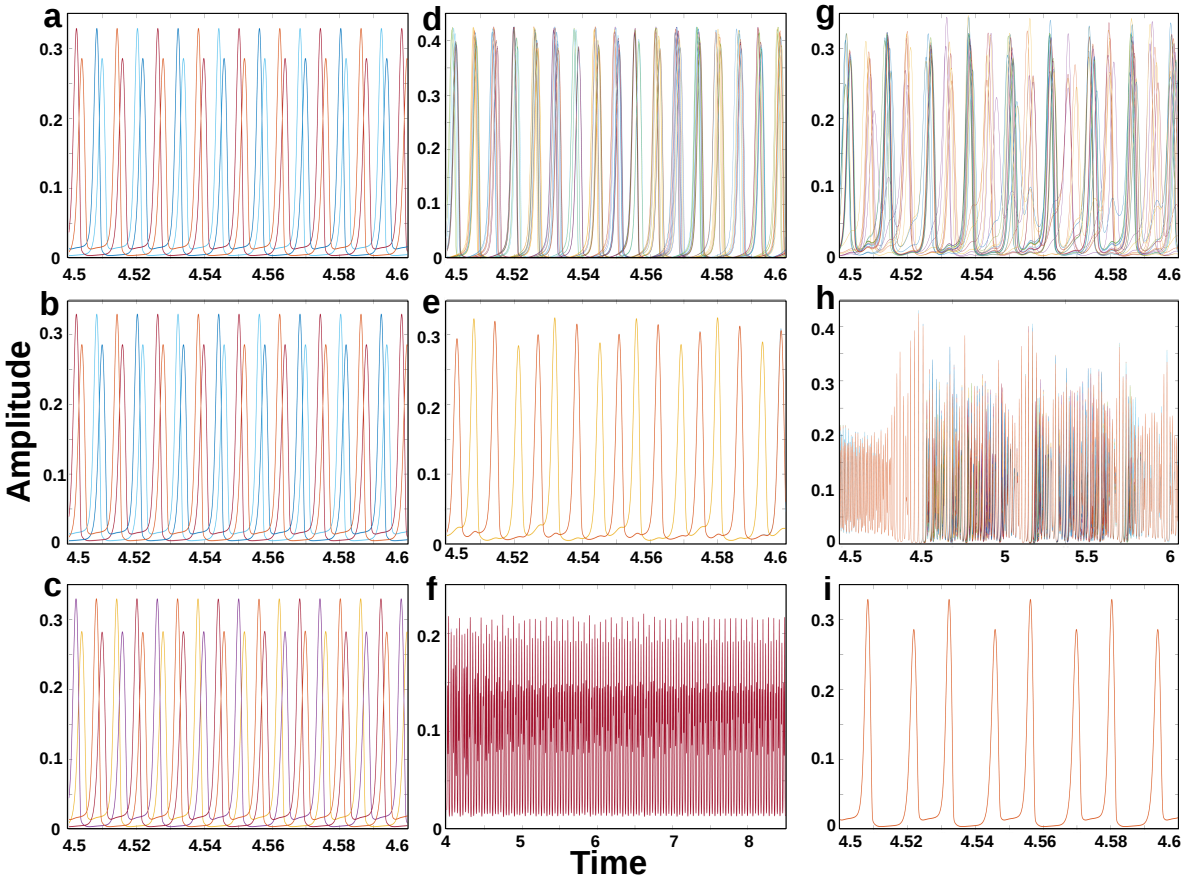

Figure 4.8: Time series of the NF−κB oscillations in global mean-field coupled oscillators for fixed TNF period T=60 and amplitude A=0.209. The time is in units of 1000 seconds. a) Splay state for zero coupling, b) Splay state for K=0.001, c) Period two splay for K=0.005, d) Asynchronous chaotic oscillations for K=0.005, e) Two groups of synchronised period-2 oscillators for K=0.05, f) Synchronised chaotic oscillators for K=0.05, g) Non-synchronised chaotic oscillations for K=0.05, h) Partially synchronised chaotic oscillations for K=0.5, and i) Fully synchronised period 2 oscillations for K=2. Note that the x-axes in panels f) and h) are different from the others.

population behaviour is same as before a 4 spliced synchronization and no new peaks are created (Fig. 4.7(b) and 4.8(b)). At $K = 0.005$ the population showed a chimera state (Fig 4.7(c)). The populations is broken into two groups. One is represented by splay states of period two attractor, with peaks at around 0.33 and 0.29 (Fig. 4.8(c)). The other group is represented by chaotic orbits with mean peak at around 0.4 (Fig. 4.8(d)). At $K = 0.05$ coupling the population is organised into 4 groups (Fig. 4.7(d)). First two are synchronised to two different period two attractors (Fig. 4.8(e)). The third and fourth groups are highly chaotic, with oscillators in third group being highly synchronised (Fig. 4.8(f)) and fourth are not synchronised (Fig. 4.8(g)). Further increase in coupling led to more synchronised population but with highly chaotic orbits (at $K = 0.1$ and 0.5) (Fig. 4.7(e) and Fig. 4.8(h)). For even larger coupling, the entire population synchronises to period-2 oscillations (Fig. 4.7(f) and Fig. 4.8(i)).

4.4 Discussion and summary

In the present work I have examined the collective dynamics that emerge through the interaction of networks of the important transcription factor, NF$-\kappa$B. The fact that the dynamics of NF$-\kappa$B in a single cell is oscillatory offers the possibility that it leads to differential gene expression [37, 188, 197, 204]. When in addition, there is bistability (or multistability), there is the possibility of dynamical switching; an effect that can arise is noise-induced hopping that can help a cell switch between high and low production states for defined genes, namely multiplexing [37]. Noise can lead to an incoherent population, as has been experimentally observed [37, 205].

In addition to states of global synchrony, the present network shows complex organization: an ensemble of cells can separate into clusters, each of which is individually synchronized. There can also be splay states as well as chimeras, namely the coexistence of coherent and incoherent populations of oscillators. In the present system, there are two types of coupling: all cells are driven externally by the periodic variation of a key enzyme, TNF, and there can be a global mean field formed by the NF$-\kappa$B in all cells. In all these cases, I find that the basins of different attractors have a complex geometry, and while it is difficult to establish that they are riddled, the fractality of the basin boundaries would indicate that the final state is highly sensitive to initial conditions. By measuring the uncertainty coefficient of the basins, I find that the amplitude and period of TNF variation play an important role in determining the global dynamics. An increase in the amplitude increase the basin uncertainty, suggesting that the system sensitivity to noise will be amplified with increasing amplitude [37, 199].

I conclude with some observations regarding the possible role of dynamical chimeras

in biological systems. As has been pointed out by Pisarchik and Feudel [34] and Pecora and co-workers [187] among others, multistability in the dynamics may be an essential requirement in biological systems since this forms the basis of many switches [206, 207]. Dynamical symmetries - such as the creation of global sychrony - and the breaking of such symmetries are both of fundamental importance in biology, especially in qualitative cellular transitions and decision-making [208]. Symmetry breaking is essential during development, immune response, hormesis etc. Dynamical chimeras, with both coherent and incoherent populations provide such an opportunity for cellular oscillators to break an existing symmetry, and this could play a role during inflammation and tissue repair, when both temporal and spatial synchrony are envisaged in a cellular population [209, 210]. Here multistability can play a crucial role.

CHAPTER 5

Eukaryotic genome expansion: A consequence of asymmetric opportunity costs?

Data from the large number of organisms that have been sequenced so far suggests that the overwhelming majority of non nucleated organisms—bacteria and archaea—have genomes that are *shorter* than 10 Mbp, with fewer than ten thousand genes. Eukaryotic genomes on the other hand, are inevitably longer than 10 Mbp and their gene numbers range from a few thousand to tens of thousands. Since genome length and gene number are quantitative proxies for genomic complexity, the observation of almost phase transition-like behaviour between prokaryotic and eukaryotic genomes is of interest. Here I examine the distribution of genome sizes as well as genome composition and discuss mechanisms that could underlie the fairly sharp bound that separates prokaryotic and eukaryotic genomic size distributions. I hypothesise the advantage of having functional endosymbionts with their autonomous genome has helped proto-eukaryotes (proto-nucleus) to effectively *outsource* the genes involved in energy production and enriched it self with regulatory genes. This allowed pro-nucleus to bypass the constraints on genome expansion, and with subsequent genome reorganisation, has enabled significantly longer genomes in eukaryotic nucleus than in prokaryotes.

5.1 Introduction

A subject of continuing interest in comparative and evolutionary genomics is the prokaryote to eukaryote transition. Like the related matter of the origin of life itself, this major

transition [211] appears to have occurred only once, but the physical reasons underlying it are not very clear. There is considerable evidence and support for a single origin of life [212]. The genomes of contemporary organisms must have therefore evolved from an initial genome that was most probably fairly small and of low complexity. Similarly, the size and complexity both have increased along the evolutionary path from prokaryotes to eukaryotes [213, 214], but our understanding of the relationship between the size of the genome and its complexity remains incomplete, leaving the question of *why* this transition occurred largely unanswered.

The present chapter is concerned with the question of whether there is a natural limit to the size of a genome (both prokaryote and eukaryote), similar to the observation made in another biological context by Haldane [215], that there is an optimal size to every type of animal, and that significant change in size was inevitably accompanied by a change in form. Although made in the context of physical size and dimensions [40], such "size optimality" arguments appear to have a wide range of applicability [215, 216, 217], and our focus here is on asking if such considerations also apply to the genome of living organisms, and if so, what separates prokaryote genomes from eukaryotes?

I first compare the distributions of number of genes and genome lengths across essentially all available prokaryotic, archaeal and eukaryotic organisms. These show a fairly sharp separation between prokaryotes and eukaryotes on the basis of genome length and to a lesser extent on gene numbers. Prokaryotes limited their gene length to 10 Mbp and gene numbers to ten thousand, while eukaryotes have exceeded these limits (see Section 3. below). This observation has been noted for some time now [213, 218, 219].

I further compare the trends in genome content with gene numbers in prokaryotes and eukaryotes. The genes on a genome have been functionally classified into COGs (clusters of orthologous genes) that help in functional and comparative evolutionary studies [220]. In earlier work it has been noted that the number of genes in particular functional categories have a power–law dependence on the total gene number. The constraint on the prokaryotic genome length and gene numbers [221, 222, 223, 224] was thus seen as a result of this growth law, namely that prokaryotic genomes could not support the proliferation of particular types of genes beyond a limit. I extend this analysis to available eukaryotic genome data, and find that many categories of genes also have a similar power–law growth with total gene number with few categories like transcription and translation differing considerably. For transcription genes in eukaryote genomes the exponent is below 1 (0.75) while for prokaryotes the exponent is larger than 1 (1.4), indicating that the differential proliferation of such genes might constrain the expansion of prokaryotic genomes.

Energy utilising genes, whose proteins participate in energy intensive functions of

secondary metabolism, or lipid and carbohydrate metabolism or translation can be contrasted with genes involved in energy production. When the size of the genome increases, the ratio of genes in these two categories reduces marginally for both eukaryotes and prokaryotes. This suggests that similar constraints operate in both these cases.

Lane and Martin [219] have recently drawn attention to sub-cellular energetics and proposed that the genome size expansion was facilitated in eukaryotes due to the excess energy available per gene even though the energy per mass unit was comparable between eukaryotes and prokaryotes. The mitochondrion, viewed as a "biophysical invention" [219] increased the energy producing membrane by many orders of magnitude. In prokaryotes, the energy production is associated with cell membrane which is smooth. Thus energy production scales with surface area, namely as volume to the 2/3rd power. As I estimate, in Section 3.3 below, the eukaryote cell has between 1 and 8 times greater power per gene per cell and between 1 to 3.3 times lower energy demand than the correspondingly sized prokaryote with similar physical and genomic dimensions. This would suggest a somewhat limited role of mitochondria in providing an innovative strategy that can increase the total energy availability.

In the early stages of the evolution of eukaryotes, namely after endosymbiosis (see Fig. 5.1) genes within the proto-nucleus and the proto-mitochondrion might have had very similar size dependence (in particular as seen in the scaling laws of gene number versus size) that would effectively constrain the expansion of genome. Assuming that the proto-nuclear genome was larger than the proto-mitochondrion genome, the opportunity cost of transcription would be smaller for the nuclear genome (power law scaling exponent greater than 1). This would lead to an enrichment genes involved in transcription in proto-nuclear genome, whereas other functional genes like energy genes can be enriched in the proto-mitochondrion, without being constrained by the scaling laws. The proto-mitochondrial genome can therefore increase the gene expression for energy genes by increasing the genome copies (see the argument in [225]). This drastically reduces the burden of carrying regulatory genes, in comparison to the proto-nuclear genome. The proto-nucleus can now increase the protein production many-fold due to increased transcription, while the mitochondria can proved the required increase in energy. Being delinked from the scaling constraints, the large exodus of proto-mitochondrial genes into the nuclear genome can restructure the proto-nuclear genome which can now expand significantly.

In light of currently available data, in the present work I re-examine genome content scaling laws and show that both eukaryote and prokaryotic genomes are constrained by similar growth curves, except in the case of genes for transcription, where there is a striking difference. In the absence of any compositional difference, the proto-nuclear genome might well have had similar scaling constraints as the prokaryote genome which

would have led to similar constraints on genome size. (All else being equal, the constraints faced in the eukaryotic cell are similar to those in the case of prokaryotes.) It therefore seems an attractive hypothesis that the relegation of energy genes to proto-mitochondria, with a concommitant reduction in the regulatory overhead and consequent enrichment of genes of other functional categories in the proto-nucleus allowed the nuclear genome to grow in size. Thus the evolution of mitochondria can also be seen as an evolutionary necessity that permitted the increase in genome size in eukaryotic cells, in addition to being an "invention" that increased the efficiency of energy production.

5.2 Materials and Methods

The genome size (length and gene number) data for bacteria, archaea and eukaryotes is obtained from the NCBI genome database [64]. The COG (Clusters of Orthologous Groups) composition of each genome is analysed using data acquired from Integrated Microbial Genomes database [65]. The data is manually curated to remove multiple representations of a given species (only a single representative genome is used in the present analysis). In total the data is composed of 3259 Bacterial, 182 Eukaryotic, 316 Archaeal genomes and 289 Bacterial plasmids.

Since the proportion of genes that are classified in COGs varies with each genome, I have normalised the numbers of genes in each functional category to its genome size by multiplying by the ratio of total gene number to the total number of COG classified genes. The number of genes for each functional category of a genome obtained in this manner is used in analysing the variation as a function of the genome size (Fig. 5.4). These data are fit to power laws using standard linear regression. The slope and intercept (pre-factor) from linear regression of data shown in Fig. 5.4 are represented in Fig. 5.5, S3, they have to be read together. The number of energy genes in an expanded prokaryote (for Fig. S4) is estimated from the linear fit of available data (from Fig. 5.4).

5.3 Analysis

5.3.1 Genome length and gene number distributions

Shown in Fig. 5.2 are two ogives for prokaryotic and eukaryotic genomes. The abscissa indicates the genome length (in the cumulative normalised frequencies, or ogive) of the number of prokaryotic genomes of size below a given length, and the complement of this curve for eukaryotic genomes, namely the fraction of such genomes above the given length. Note also that the abscissa is in logarithmic scale and that the ogives cross at $10^{6.9}$, namely a length of 8.13 Mbp. This figure is based on currently available data from

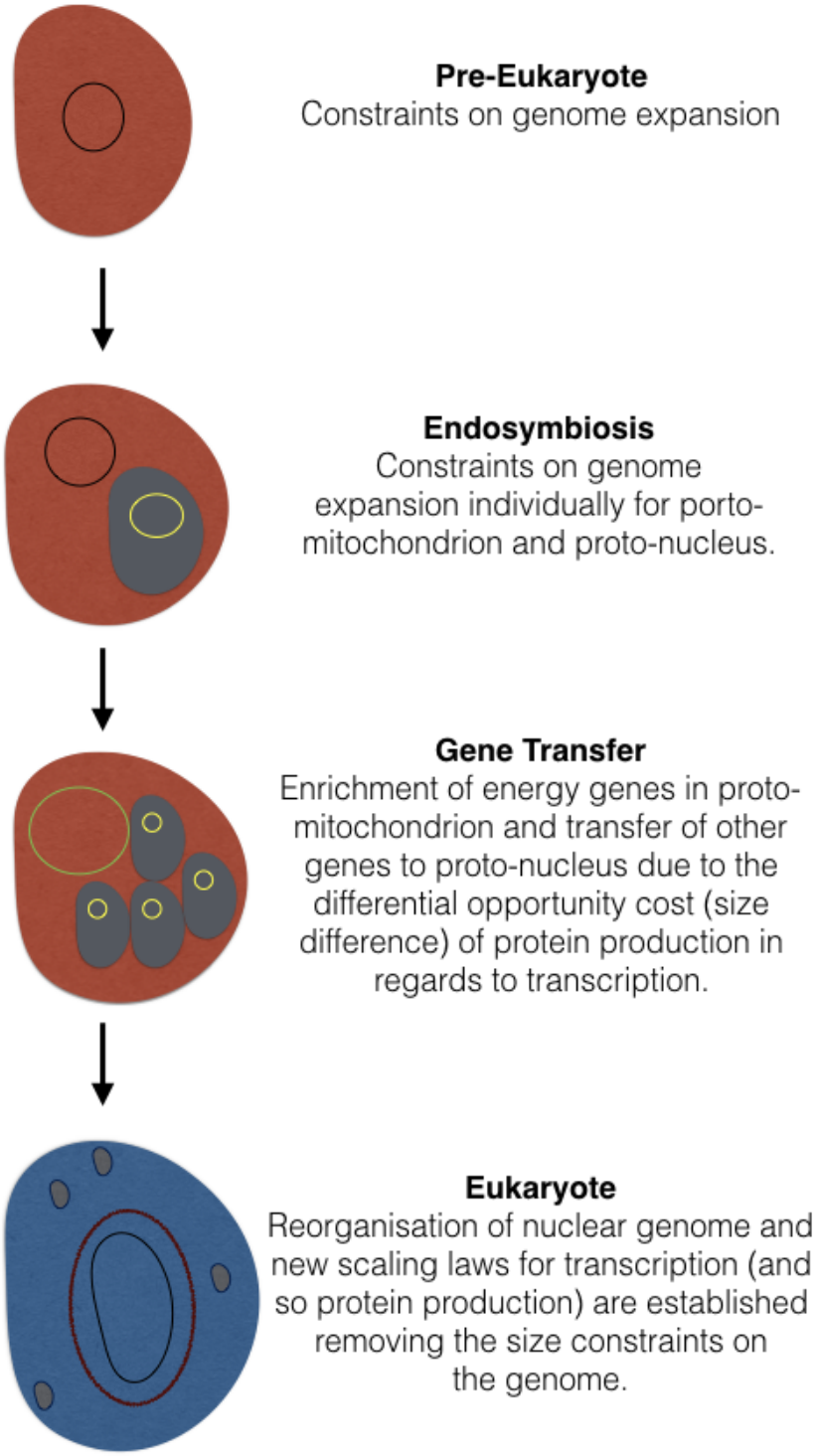
Pre-Eukaryote
Constraints on genome expansion
Endosymbiosis
Constraints on genome expansion individually for porto-mitochondrion and proto-nucleus.
Gene Transfer
Enrichment of energy genes in proto-mitochondrion and transfer of other genes to proto-nucleus due to the differential opportunity cost (size difference) of protein production in regards to transcription.
Eukaryote
Reorganisation of nuclear genome and new scaling laws for transcription (and so protein production) are established removing the size constraints on the genome.

Figure 5.1:

the genomes sequenced so far, and it seems unlikely that the basic shape will change drastically with time. One can also see that the boundary is in fact sharp. From the crossing point which occurs at $L \approx 8.1$Mbp, less than 4% of prokaryotes exceed the length and less than 4% of eukaryotes are below the length.

Similar data for the distribution of the total gene numbers show that prokaryote genomes are limited to about 10^4 genes whereas eukaryotes can have up to ten times that number, although several have fewer than 10000 genes. These bounds have been noted earlier [214, 226, 227], and the presently available data is summarised in Fig. 5.3.

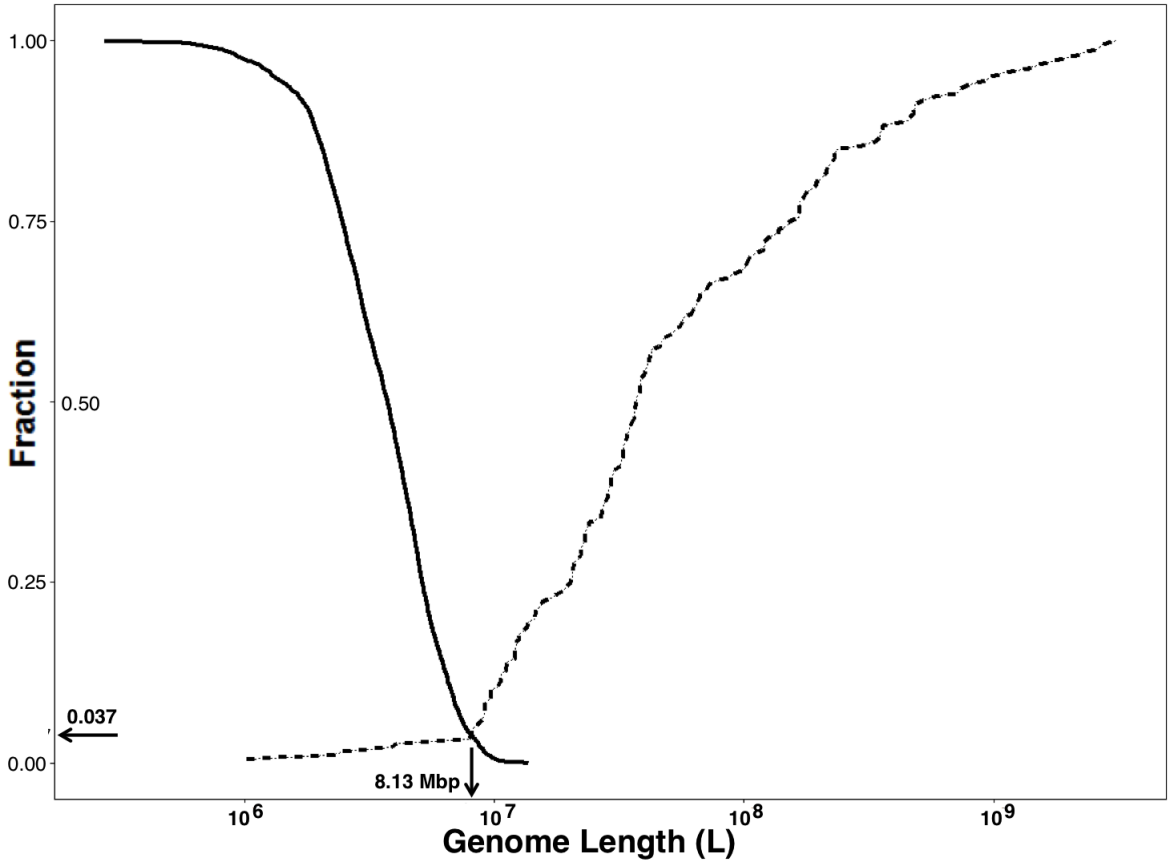

Figure 5.2: Shown with solid curve is f_p the fraction of prokaryotic genomes *longer* than a given length L, while the dashed curve is f_e the fraction of eukaryotic genomes *shorter* than L. The curves cross when $L = 10^{6.91} = 8.1$ Mbp, $f_p = f_e = 0.037$. Note the logarithmic scale on the abscissa and I considered total genome length. If the data for eukaryotes is segregated by chromosome, then the separation between curves (see Fig.S1, chapter appendix) is not strong.

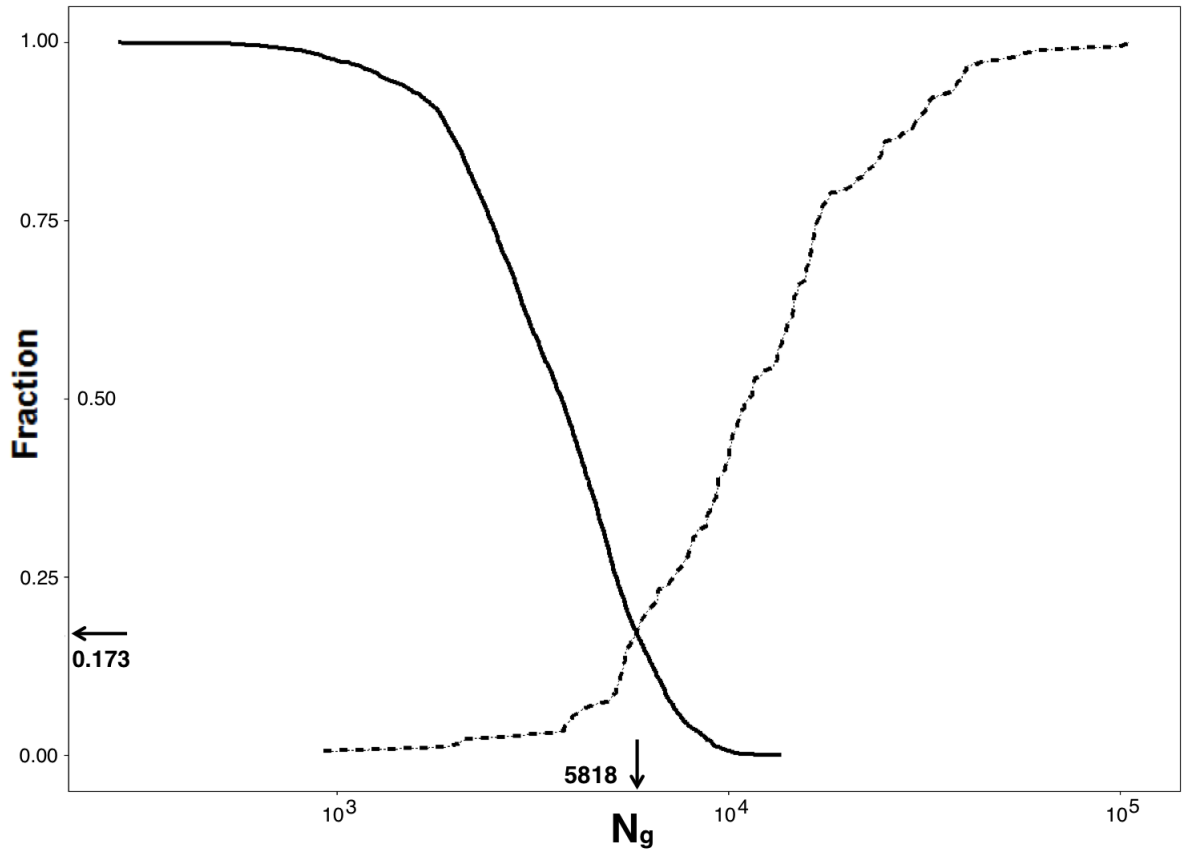

Figure 5.3: Shown with solid curve is f_p the fraction of prokaryotic genomes *longer* than a given gene number N_g, while the dashed curve is f_e the fraction of eukaryotic genomes *shorter* than N_g. The curves cross when $N_g = 10^{3.764} = 5818, f_p = f_e = 0.173$. Note the logarithmic scale on the abscissa.

5.3.2 Genome content

Following Konstantinidis and Tiedje (2010) [228] I segregate the genes in each genome into COG functional categories, and examine the variation in the number of genes in a given COG as a function of the total number of genes in the genome. Attention is restricted to those COGs that constitute at least 3 % of the genome (these are shown in their entirety in Fig. S2, chapter appendix). A representative subset is shown in Fig. 5.4 where data for both eukaryotes and prokaryotes are presented together. The scaling exponents of all these categories for both prokaryotic genome and eukaryotic genome is shown in Fig. 5.5. For a power law $x = ay^b$, a is the pre-factor and b is the scaling exponent. Within a simple evolutionary model [221] the scaling exponent corresponds to the ratio between selection of duplicated genes in the COG compared to selection of duplicated genes in general.

In most COGs the number of genes increase as a power–law, while some others (data not shown) show saturation. Konstantinidis and Tiedje [228] suggest that power–law growth of genes in these categories (specifically transcription regulation) is a major reason why prokaryotic genomes are limited in size; see also [222]. In most COGs there are no major differences evident in the scaling behaviour in eukaryote or prokaryote genomes, except for those genes involved in transcription (0.75 and 1.4) and translation (0.7 and 0.32).

The former category genes increase more slowly with genome length in eukaryotes in comparison to prokaryotes, while in the latter category, the eukaryotic genes increase somewhat faster. However, the percentage of genes for translation progressively reduces for both eukaryotes and prokaryotes (scaling exponents less than 1), and therefore this difference may have had no significant role to play in constraining the evolution of large genomes. On the other hand, the large difference between the growth scaling exponents for transcription genes (including those involved in regulation of transcription) between prokaryotes and eukaryotes is significant.

The data suggests that for transcription genes, eukaryotes and prokaryote have contrasting behaviours: with a scaling exponent of 1.4, the transcription genes of prokaryotes have increased selection compared to the average, whereas with an exponent of 0.75, transcription genes are selected lower than the average gene in eukaryotes. The difference between eukaryotes and prokaryotes is largest for this category (in comparison to all others) and is suggestive of changed genome architecture in regards both to transcription as well as to the differential constraints on the genome size between prokaryotes and eukaryotes.

Since energetics is cited as a prescriptive reason for evolution of large genomes [219], the scaling exponent for energy related genes is of interest. There is only a small difference between these in eukaryotes and prokaryotes. But prokaryotes have scaling exponent slightly greater than 1 (1.08) where as eukaryotes have it very near to 1 (0.96). Which means in eukaryotes they enrich on par with the average category in the genome, where as in prokaryotes they enrich slightly more than average. I clubbed together the COGs of genes responsible for Amino acid, Lipid, carbohydrate Nucleotide, Coenzyme and Inorganic Ion transport and metabolism, translation and secondary metabolism [220], namely all the energy utilising genes in order to compare with the enrichment of energy producing genes. The ratio between these two categories was calculated for all organisms under study and this is plotted in Fig. 5.6. Both prokaryotes and eukaryotes show a slight negative correlation but the difference between categories is low and has a large scatter.

If one assumes that a prokaryote genome expanded so that the number of genes was

the same as in a eukaryote, it is possible to estimate the number of energy genes in such a hypothetical prokaryote using the scaling laws. Such estimate can then be compared with the corresponding eukaryotic case. There is a good correlation over a considerable size range and only deviating for examples of very large genome sizes (See chapter appendix Fig. S4). These observations suggest that the enrichment of energy genes can be a constraint on prokaryote genome expansion but is a smaller constraint compared to the enrichment of transcription regulation genes.

Considering values for all other categories, bacteria considerably enriches more of signal transduction (1.78 vs 1.26), secondary metabolites biosynthesis, transport and catabolism genes (1.8 vs 1.51) compared with eukaryotes and eukaryotes enriches considerably more of defence (1.34 vs 0.97) and cell cycle cycle control, cell division, chromosome partitioning genes (0.8 vs 0.6). In all these categories the direction of enrichment that is whether the category enriches more or less than general genes (scaling exponent greater than 1 or lesser than 1), do not differ between eukaryotes and prokaryotes.

5.3.3 Energetic differences

As discussed above, the difference between the scaling exponents for the different COG categories in bacteria and eukaryote genomes for either energy production or energy utilisation genes is small. In order to contrast the gross energetics of the prokaryotic and eukaryotic cell (as in Lane (2011) [226]), I consider them to be spherical with radii r_p and r_e respectively.

Estimates for the mean weight of "typical" cells, as well as their energy efficiency and the power per cell are summarised in Table 5.1 (data from Lane (2010) [219]) The ratio of the radii is approximately

$$\frac{r_e}{r_p} = \sqrt[3]{\frac{40100}{2.6}} \approx 25. \tag{5.1}$$

Since the power within the cell is proportional to the surface area, if a prokaryotic cell were to be expanded by a factor of 25, the power would increase by a factor of 625, namely

$$P_2 = P_1 \left(\frac{r_e}{r_p}\right)^2 = 0.49 \times 25^2 \approx 310\text{pW/cell}. \tag{5.2}$$

In order to estimate the ratio of power per Mbp per cell for eukaryotes to the expanded prokaryote, I proceed as follows. Consider that the number of genes in the expanded prokaryote are increased to match the number q as in a typical protozoa. The ratio of metabolic power per Mbp per cell is

$$\frac{2286/q}{310/q} \approx 8. \tag{5.3}$$

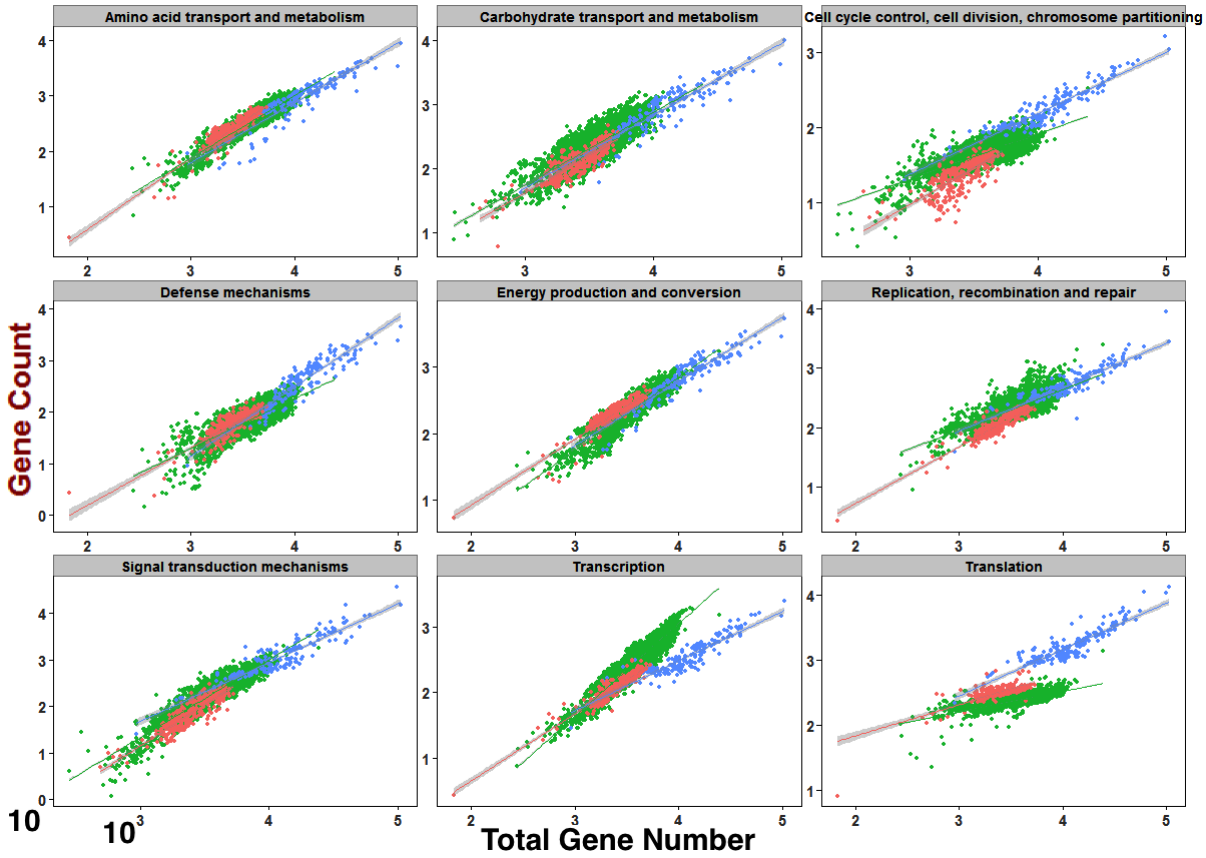

Figure 5.4: Dependence of the number of genes in each functional category on the size of the genome. Each graph corresponds to a different functional category. Data for archeal genomes is in red, for bacterial genomes in green, and for eukaryotes in blue. Both abscissa and ordinates in each of the graphs is in a logarithmic scale.

	Prokaryote	Eukaryote
Mean Weight	2.6×10^{-12} g	40100×10^{-12} g
Efficiency	0.19 pW/g	0.06 pW/g
Power per cell	0.49 pW/cell	2286 pW/cell

Table 5.1: Summary of energy efficiency from Lane (2011)[219].

Assuming that the doubling time for the expanded prokaryote and a similar eukaryote is the same, the ratio of energy per Mbp per cell will be equal to that of power. Were the surface of the cell to be treated as a disordered membrane, then the area would scale as some exponent greater than 2 with respect to the radius, giving a somewhat lower figure

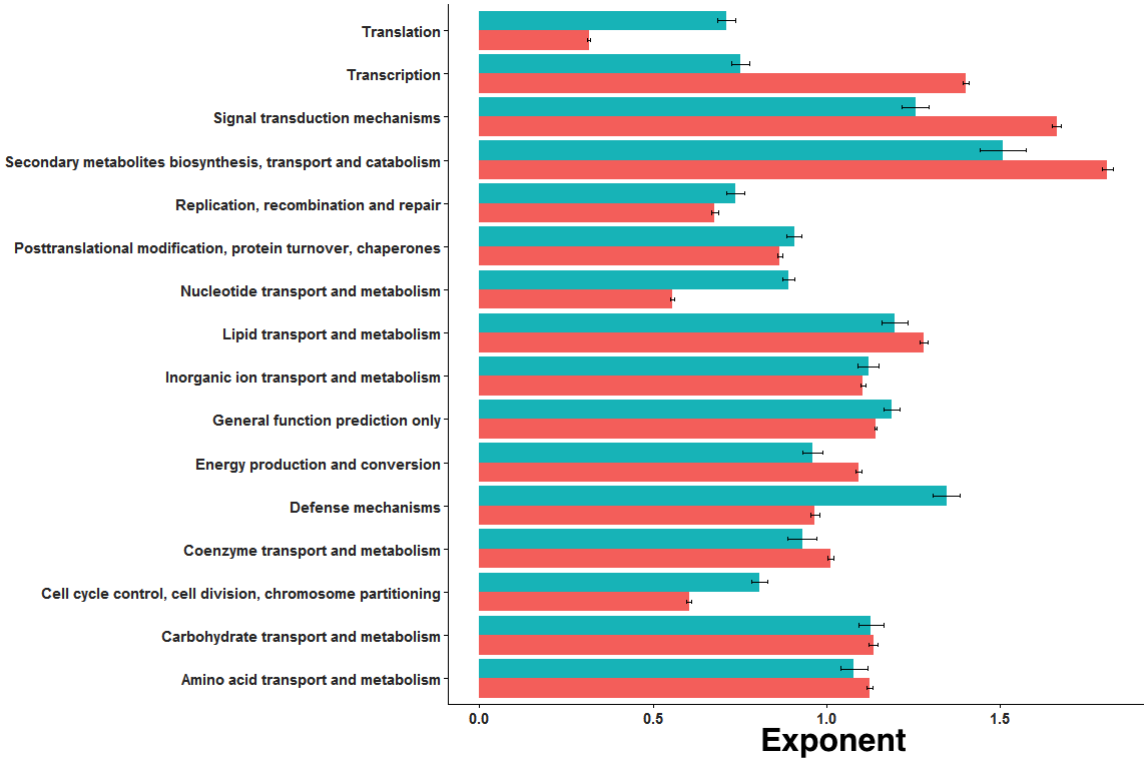

Figure 5.5: Comparison of the slopes of the power–laws in Fig. 5.4 . The data for bacteria is shown in red while that for eukaryotes is in blue.

for the ratio (at 2.4 the ratio will be 1). In terms of the energy required for translation, or for any other function that scales by volume (since this is the same for the expanded prokaryote and eukaryote) one finds that at eukaryotic sizes, prokaryotes are less energy efficient by a factor of 1 to 10. If the energy demand is considered same for eukaryote and expanded prokaryote (of same size) then energy efficiency differences are less than an order of magnitude at best.

On the other hand, the assumption that the hypothetical enlarged prokaryote that is the size of a eukaryote (both in terms of volume as well as genome size) would still have energy demand at the prokaryote rate is reasonable. For a prokaryote, the energy demand will be equal to energy produced and is 0.49 pW/cell. Considering energy demand scales with volume, the expanded prokaryote will have a demand of 25^3 as much, namely 7656 pW/cell. Similarly a eukaryote will have energy demand equal to its energy production, namely 2286 pW/cell so that the ratio of energy demand between the expanded prokaryote and eukaryote is around 3.5 times. This ratio can be increased to about 25 if one takes into consideration the difference in energy efficiency and after adjusting for energy demand between eukaryote and expanded prokaryote (see the chapter

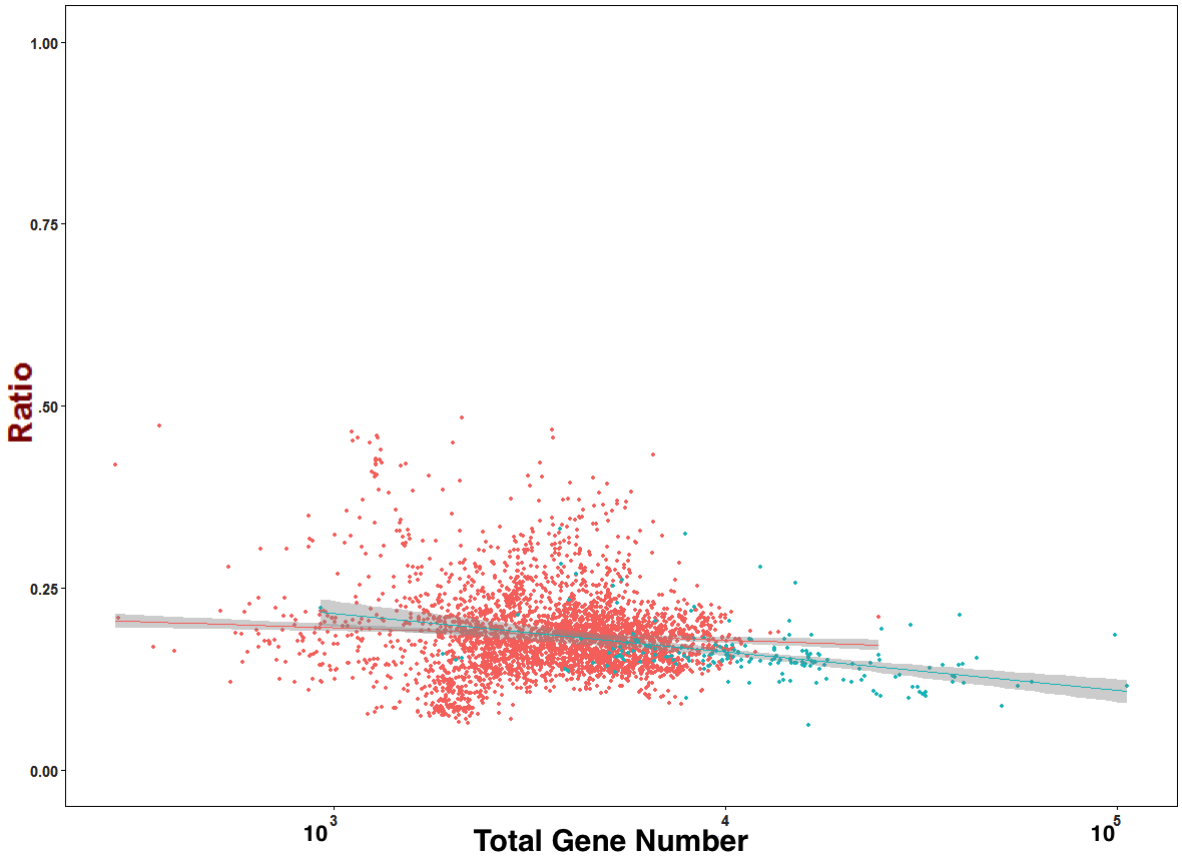

Figure 5.6: The ratio of energy production and conversion genes to energy utilisation genes as a function of the genome size. There is a negative correlation for both bacteria and eukaryotes.

appendix, in particular Fig. S4, for details). The above estimates rest on the idea [226] that the rate of protein production in the expanded prokaryote genome can increase proportionally to the cell volume.

This shows that the difference in energy production between prokaryotes and eukaryotes is not large; a hypothetical prokaryote of that had eukaryotic dimensions could generate equivalent power by scaling its surface area by a factor of 2.5, but the difficulty is in increasing the rate of protein production by 25^3 times with only a ten-fold increase in genome length and a still smaller increase in gene number. Eukaryotes could do this. Similar problem of scaling would have occurred when prokaryotes and Eukaryotes increased there genome and cell size from a simpler ancestor, and so should have a characteristic scaling factor for increase in protein production with genome size. That is a growing genome have to increase the energy production but for doing so there is a steeper problem of increasing protein production. And so the fairer question to be asked

is that, if there exists a characteristic scaling for protein production with genome size in prokaryotes and eukaryotes. If so how large a prokaryote genome can grow before it gets constrained? And how eukaryotes have done it differently?

Transcription factors regulate gene expression and it is just to assume that the number of transcription genes correlate with the quantum of protein production. The requirement of increase in protein production can now be posed as the requirement for increase in transcription genes. This suggest that, as discussed in previous works [221, 222, 223] and from our data (Fig. 5.5), the power law scaling of transcriptional regulation with an exponent greater than 1 (1.4) is at the crux of constraints on genome size in prokaryotes.

5.3.4 Telos of mitochondria and its genome

As discussed in the previous section, the power production in a typical eukaryotic cell is at most an order of magnitude of what might be expected in a prokaryotic cell of the same size. This suggests that the increase the energy production in eukaryotes provided by mitochondria may be quite limited. Furthermore, the number of energy producing genes in the eukaryotic genome is very similar to the (expanded) prokaryote, and as has been noted earlier [229] mitochondria play a role in eukaryogenesis that possibly extends beyond their obvious and well-known function in energy production.

Consider the last eukaryotic common ancestor (LECA). This organism had a proto-mitochondrion (most likely an α-proteobacteria) that would have the same scaling laws as other prokaryotes. In order to increase the genome size, LECA would have needed to increase the production of energy proteins scaling similar to volume (by a factor of 15625 if it had to grow to the size of eukaryote), which would mean an increase in the number of genes in transcription and translation functional categories in line with the prokaryotic scaling laws. This is the genomic opportunity cost of an increase in energy proteins, and thus the cost of genome expansion.

The scaling exponent for translation genes is lower than 1: as the genome grows larger the opportunity cost of protein production due to translation reduces. On the other hand, the scaling exponent for transcription genes is 1.47, and thus this increases with genome size. I assume that the genome of the proto-mitochondrion is smaller than that of LECA, and further, that the LECA genome is long enough that the reduction in translation cost can offset transcriptional cost. Then the opportunity cost of production of energy proteins is lower in the proto-mitochondrion than in the LECA genome: in this case, enrichment of energy genes in the nuclear genome would not be burdened by needing a large transcription machinery. The associated question of why chloroplasts and mitochondria contain so many genome copies has been ascribed to differences in the mechanism by which gene expression is controlled [225]. In compensation for not having

as elaborate gene control as nuclear or bacterial nuclei that can increase the expression of specific genes manyfold, mitochondria have to increase copy number to match the required rRNA levels [225].

It would have been a similar case where the required increase in proteins for energy production is achieved by increase in genome copies of protomitochondrion which is enriched with energy genes and with lower regulatory (transcription) overhead. This arrangement would help the early eukaryotes to bypass the scaling constraints on genome and would lead to a condition where the genome grows in size and with no reduction in energy per gene nor energy per volume. This larger genome can reorganise and bring together all the innovation which was not possible to be collected in a single prokaryote. With limited transcriptional regulation the increase in copy number of genome will lead to unnecessary increase in expression of all other genes genes in proto-mitochondrion. As argued [226], in order to reduce the cost, other genes will be transferred from the proto-mitochondrion to the proto-nucleus, leading to the enrichment of other functional categories in the proto-nucleus.

5.4 Summary and Discussion

Our main interest in the present work has been on whether there is a natural constraint to the size of the prokaryotic genome, and how this constraint is bypassed in eukaryotes. The eukaryotic cell is structurally highly complex, with numerous structural and constitutive differences when compared to a typical prokaryotic cell, possessing a nucleus, internal membranes, heterozygosity, an increased cell size, longer genomes and complex gene structure, a cytoskeleton, and so on. Although prokaryotes have explored several biophysical innovations individually, they have not accumulated all of these in a single instance [219].

The increase in energy available per gene is an important factor in allowing for the genome expansion. One approach here has been to examine the genome content [228] in terms of specific gene families. Since there is a correlation between the complexity of the organism and the size of its genome, the number of genes of a given functionality changes with genome length. In most cases, there is a scaling behaviour, namely a power-law dependence. An exponent that is greater than 1 indicates gene functionalities that promote larger genomes and contrariwise, exponents smaller than one indicate functionalities that are not essential to increase of genome length. Current arguments are based on the difference in energetics between prokaryotes and eukaryotes, and this has been termed the energy divide [219]. Our analysis of the presently available data from this point of view, presented in Section 3 above, seems to suggest that eukaryotes

significantly differ from prokaryotes in scaling of transcriptional genes and in their requirement of large increase of protein production. The scaling behaviour of most families of genes in both categories is the same, and the absolute energy divide appears to be a factor of ten or less in eukaryotes. I also compare the typical eukaryote genome to a hypothetical expanded prokaryotic genome, and here the difference appears marginal in respect to number of energy related genes and the ratio between energy producing and energy consuming genes.

While there are several facets to the genome expansion, I have focused here mainly on scaling and energetic aspects of the genome. The scaling laws for genes involved in transcription and the energy related genes of prokaryotes suggest that the opportunity cost, namely the number of transcription genes per energy related genes increases with the genome size. Genomic asymmetry between LECA and the protomitochondrion during eukaryogenesis will result in a asymmetric opportunity costs for energy related protein production. I hypothesise that this asymmetric opportunity cost could have led to enrichment of energy genes in protomitochondrion, and could have supplied the necessary increase of protein production for genome expansion of LECA. The increased expression by increase in genome copies of protomitochondrion would lead to over-expression of many genes. To reduce such a cost, large parts of its genome are transferred to the nuclear genome and thus helps in its remodelling, and in this sense, one may infer that the eukaryotic cell could have used mitochondrial colonies in order to increase the size and complexity of the nuclear genome.

5.5 Chapter Appendix

5.5.1 Alphabet codes for COG functional category

(B) = Chromatin structure and dynamics, (C) = Energy production and conversion, (D) = Cell cycle control, cell division, chromosome partitioning, (E) = Amino acid transport and metabolism, (F) = Nucleotide transport and metabolism, (G) = Carbohydrate transport and metabolism, (H) = Coenzyme transport and metabolism, (I) = Lipid transport and metabolism, (J) = Translation, ribosomal structure and biogenesis, (K) = Transcription, (L) = Replication, recombination and repair, (M) = Cell wall/membrane/envelope biogenesis, (N) = Cell motility, (O) = Posttranslational modification, protein turnover, chaperones, (P) = Inorganic ion transport and metabolism, (Q) = Secondary metabolites biosynthesis, transport and catabolism, (R) = General function prediction only, (S) = Function unknown, (T) = Signal transduction mechanisms, (U) = Intracellular trafficking, secretion, and vesicular transport, (V) = Defense mechanisms.

5.5.2 Supplementary Figures

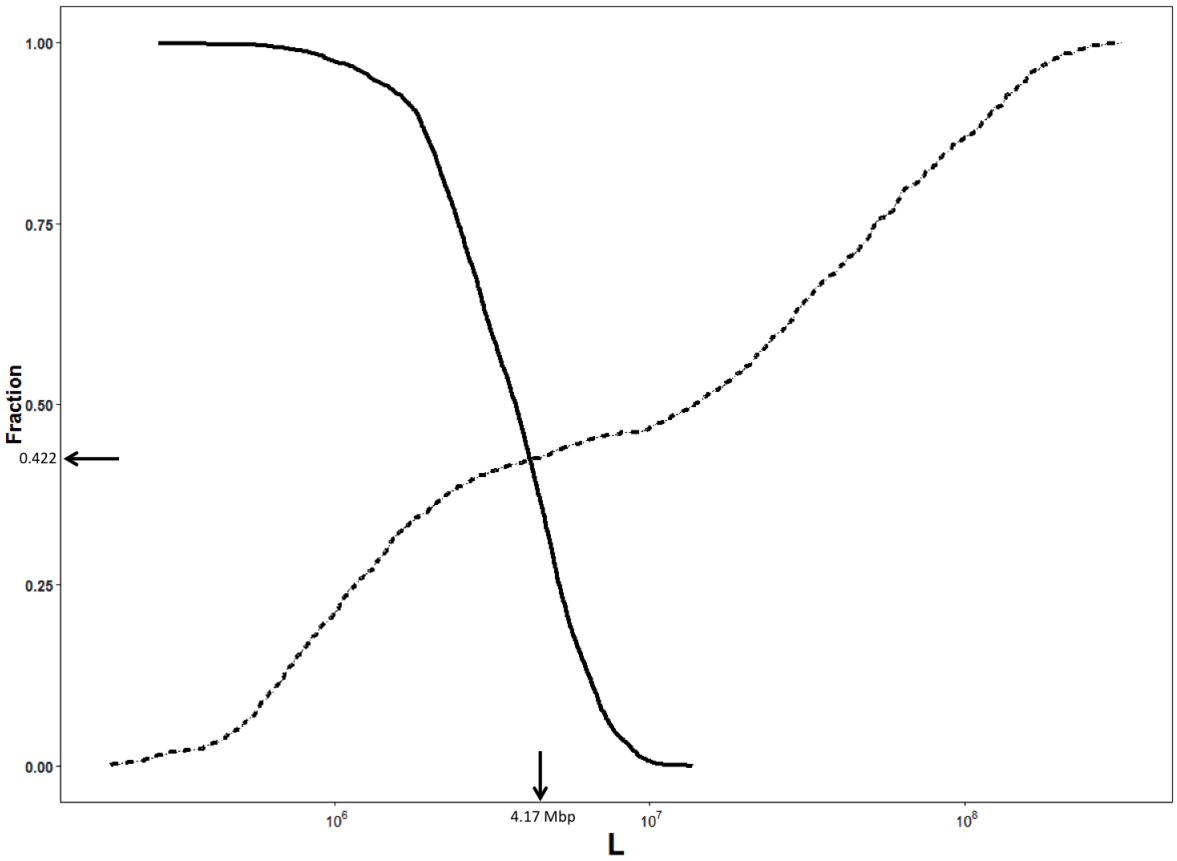

Figure 5.7: Shown with solid curve is f_p the fraction of prokaryotic genomes *longer* than a given length L, while the dashed curve is f_e the fraction of eukaryotic chromosomes *shorter* than L. The curves cross when $L = 10^{6.62} = 4.17Mbp$, $f_p = f_e = 0.422$. Note the logarithmic scale on the abscissa.

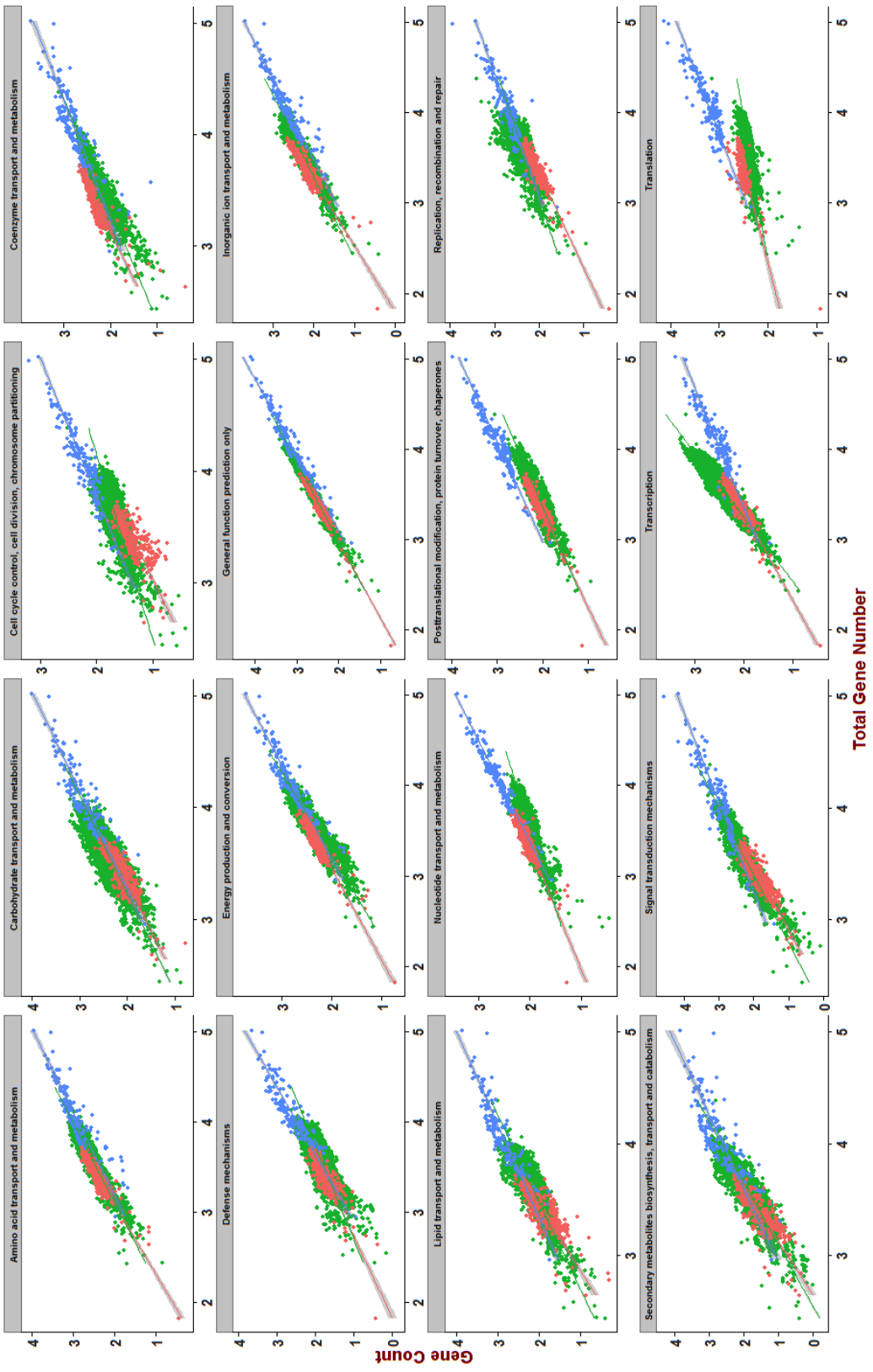

Figure 5.8: Dependence of the number of genes in each functional category on the size of the genome. Each graph corresponds to a different functional category. Data for archeal genomes is in red, for bacterial genomes in green, and for eukaryotes in blue. Both abscissa and ordinates in each of the graphs is in a logarithmic scale

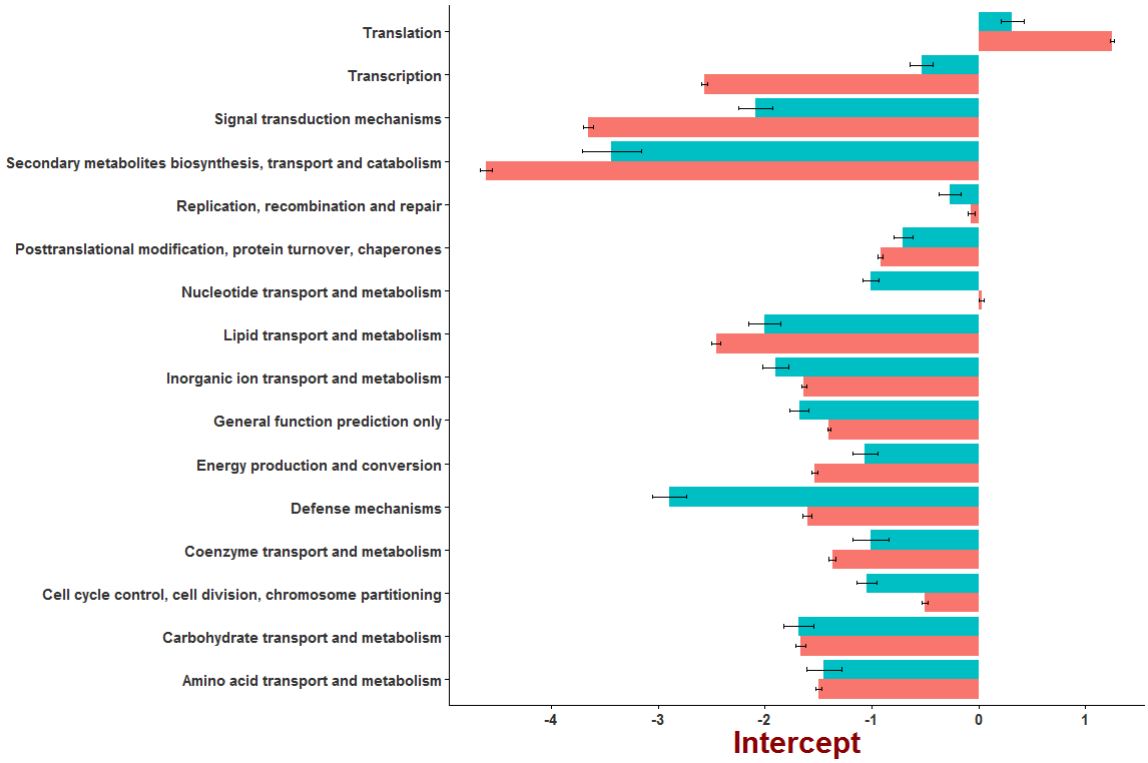

Figure 5.9: Comparison of the pre-factor of the power–laws in Fig. 3 . The data for bacteria is shown in red while that for eukaryotes is in blue.

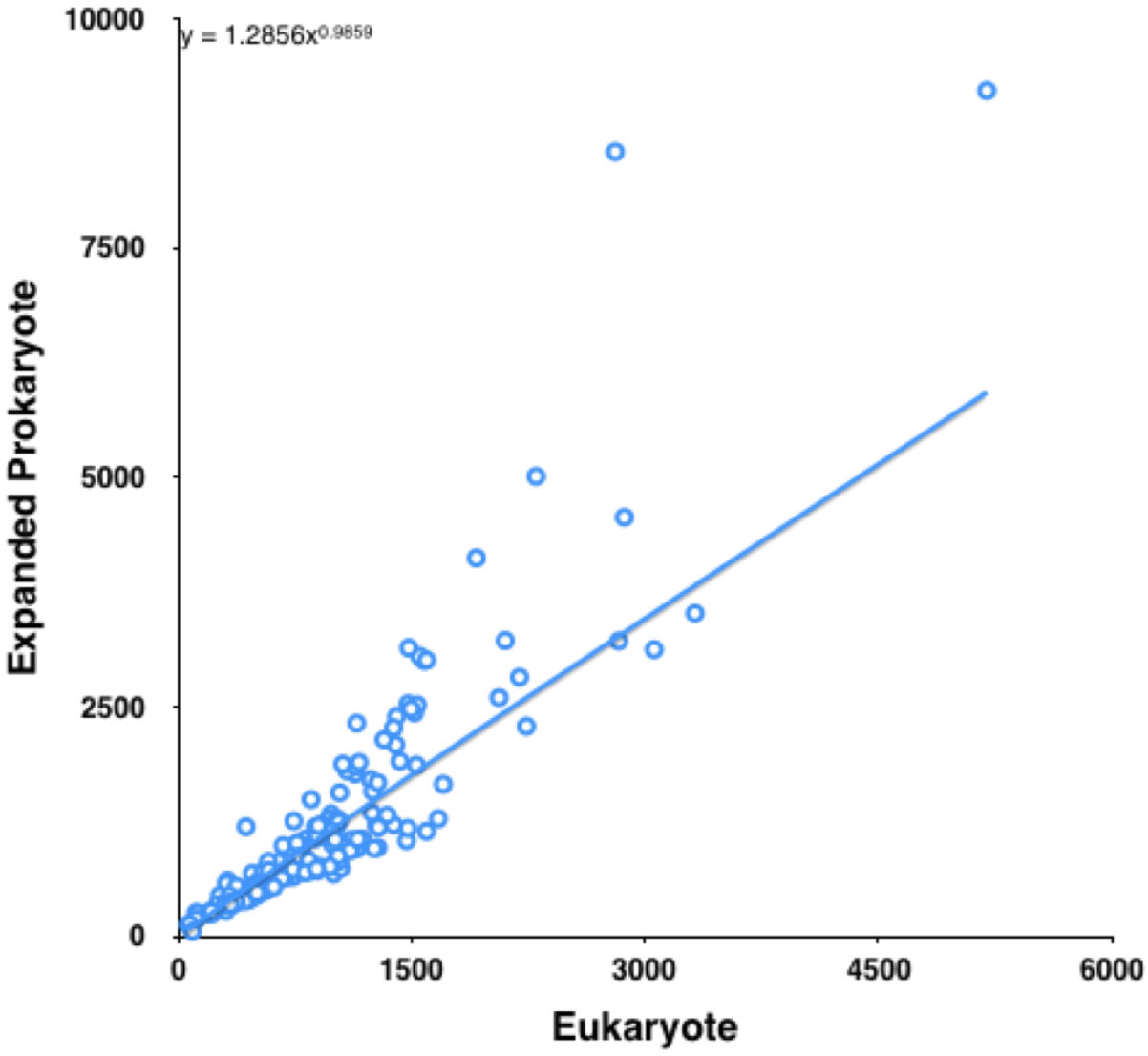

Figure 5.10: Comparison of the energy genes in the expanded prokaryote with eukaryotes.

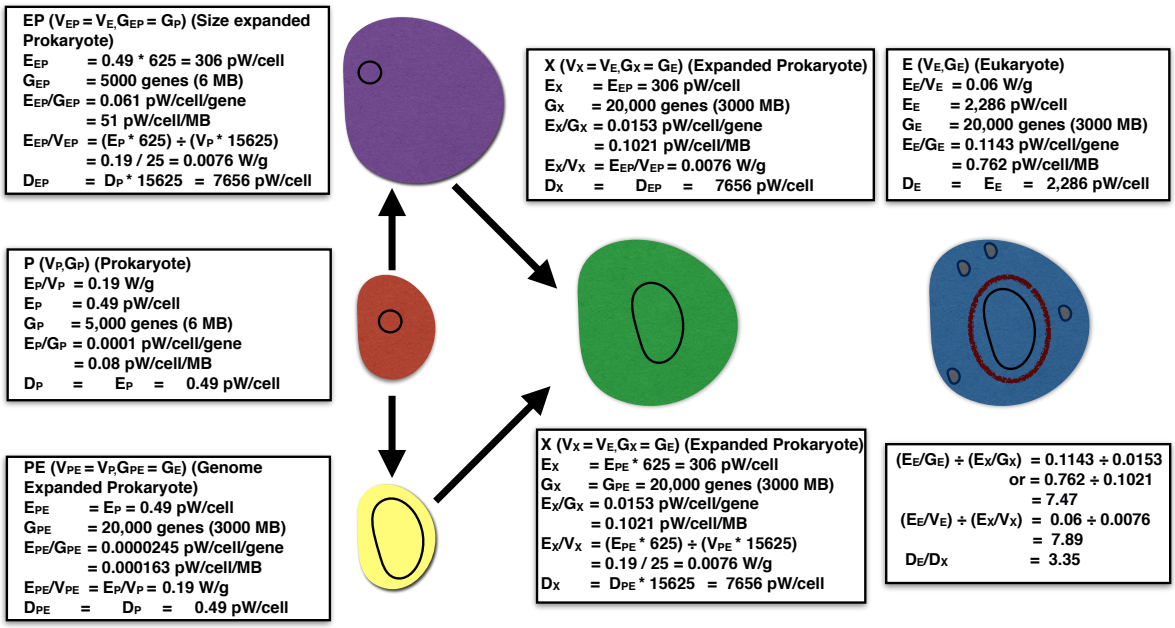

Figure 5.11: Graphical representation of the calculations for power per gene per cell of expanded prokaryote and Eukaryote. E = Energy, G = Genome Length or Gene number, V = Volume and subscripts P = Prokaryote, PE = Genome Expanded Prokaryote, EP = Volume Expanded Prokaryote, X = Prokaryote of Eukaryotic dimensions and E = Eukaryote.

CHAPTER 6

Summary and Open Problems

6.1 Thesis Summary

In this thesis I have studied a number of biological phenomena at different scales, through the formulation of appropriate models. Most biological phenomenon are in essence defined by their feature of having several levels of interaction, and are characterised by complex spatial and temporal organization. One common theme that runs through the work presented here is the emergence of novel collective behaviour. The specific problems represent biological systems at various levels, ranging from the social organization of small mammals and insects to the occurrence of synchrony and complex spatiotemporal dynamics in cellular and subcellular networks. Consequently, a variety of different modelling techniques are required to be employed. I have used mathematical and computational approaches such as evolutionary game theory, agent based modelling, dynamical systems techniques, genome analysis, branching process models, machine learning and network theory.

One of the more remarkable "laws" that apply to biological systems is that first enunciated by Kleiber [105] that relates the specific metabolic rate of an organism to its body mass. Extending over something like 20 orders of magnitude, this relationship expresses the allometric scaling principle that the metabolic rate varies as the -1/4th power of of the body mass. Thus, the life-span of an organism increases in an inversely proportional manner, namely as the 1/4th power of the body weight. There are a number of features that contribute to the longevity of organisms in addition to size, and the underlying reasons for this scaling law has been of considerable current interest [230]. One

instance where there is an anomalous departure from this law is in a population of eusocial insects, where there is a significant difference in size between the breeding (queens) and non-breeding (worker) individuals. There is an even more significant difference in the lifespans which can be as much as two orders of magnitude longer for queens as compared to workers.

I have studied the relationship between eusociality and the issue of long-lived queens, both of which appear to be connected, in Chapter 2. Developing upon an evolutionary dynamics model with age and population structure, my aim has been to incorporate standard natural selection and compare and compete populations with various population structured and differing ageing strategies. I showed that the solitary populations selects short lifespans and eusocial population structure selects queens with long lifespans. The long-lived queens in turn supports the evolution of eusociality. The other important result, coinciding with empirical observations, is that with the same external death rate, when competed, the polygynous eusocial populations are selected for a comparatively shorter lifespan than are monogynous eusocial populations.

In Chapter 3, I explored the benefits of food-sharing behaviour in a social colony of vampire bats. Cooperative behaviour provides individuals of the colony with both direct and indirect benefits. Direct benefits are comparatively easy to observe and quantify, while indirect benefits such as by-product group benefits are occur at a group level and are difficult to observe. Our interest was in whether the cooperative behaviour of food sharing, expressed in a population, will show any group benefits. I implemented the cooperative behaviour in an agent-based model, and our simulations showed that group benefits become substantial only when the majority of the population exhibits resource sharing.

The collective dynamical states of an ensemble of intracellular NfκB molecular oscillators was studied in Chapter 4, using dynamical systems theory. Synchronisation is one of the most common collective phenomena found in coupled oscillator systems in nature. A host of other collective dynamical phenomena such as splay states, cluster synchronisation, chaos and chimeras are also known. When the oscillators are coupled by an external TNF (tumour necrosis factor) oscillator, the ensemble shows all the above mentioned dynamical behaviours at different parameter values with both local and global coupling.

In Chapter 5, I have presented an analysis of the distribution of genome sizes across prokaryotes and eukaryotes. This analysis, enabled by the widespread sequencing of the genomes of a large number of bacterial and eukaryotic species, shows that there is a sharp boundary between unicellular and multicellular organisms. The genome size in prokaryotes is limited to 10 Mbp and ten thousand genes, while such a limit does not exist

for eukaryotes. Common limitations to growth involve the energetics, kinetics, material constraints, and so on. By analysing the manner in which the number of genes associated with transcription regulation increase with genome size, I showed that this might have a role in limiting the expansion of a bacterial genomes. Given this and related observations I developed a hypothesis. that the opportunity cost of energy and transcriptional genes differs between the last eukaryotic common ancestor and the protomitochondrion. The use of mitochondria as "colonies" by cellular nuclei is discussed as a possible reason that enables eukaryotes to bypass the scaling constraints acting in the ancestral genomes.

In this final Chapter, I further discuss two open problems that I have been working on, and present some preliminary results. These results are interesting and utilize other computational approaches for understanding collective behaviour in biological systems.

6.2 Heterogeneous differentiation of CD8+ T cells

The immune system harbours a large number of CD8+ T cells that are an important part of adaptive immunity [231]. For any given antigen, there are between 10^2-10^3 naive cells, and upon infection, if the antigen is recognized, the CD8+ cell specific to that antigen will increase in population by a factor of two or more. The resultant effector cells clear the infection [231]. Subsequent to pathogen clearance, a few so–called "memory cells" are left behind; these survive for much longer periods after the infection [232] and help in more rapid immune response in case of a secondary infection.

This expansion of the CD8+ T cells is highly reproducible for a specific infection. The *in vivo* lineage tracing of CD8+ T cells has been studied through two different experimental techniques, DNA barcoding [233] and immunological coding [234]. Observations from these different studies suggest that expansion leads to heterogeneity in both clonal family sizes and marker distribution. Buchholz *et al.* [234] compared a number of different models and concluded that the the data best supported a linear differentiation pathway: naive T cell $\rightarrow$ TCM $\rightarrow$ TEMp $\rightarrow$ TEF, with proliferation rate increasing through the pathway. On the other hand Gerlach *et al.* [233] conclude that asymmetric division by itself cannot explain the the experimentally observed disparity between individual T cell families, and further, they note there is a strong heterogeneity at the single-cell level indicating that T cell response can be highly variable.

In order to reconcile these contradictory models of the differentiation pathway, a new approach is required [235]. Each type of CD8+ T cell is experimentally identified through the markers they carry. Hence a population model that is devoid of the cellular types, but has only stochastically distributed markers is suggested. I propose a branching process for the proliferation of cells, with stochastically distributing inheritance factors

that can approximate the observed behaviour. The branching process is augmented with the modified cyton model that has been defined in the context of B-cell expansion [236]. Three markers are assumed, which are stochastically inherited at each cell division, and have different rates of accumulation. The model replicates the experimental observations qualitatively.

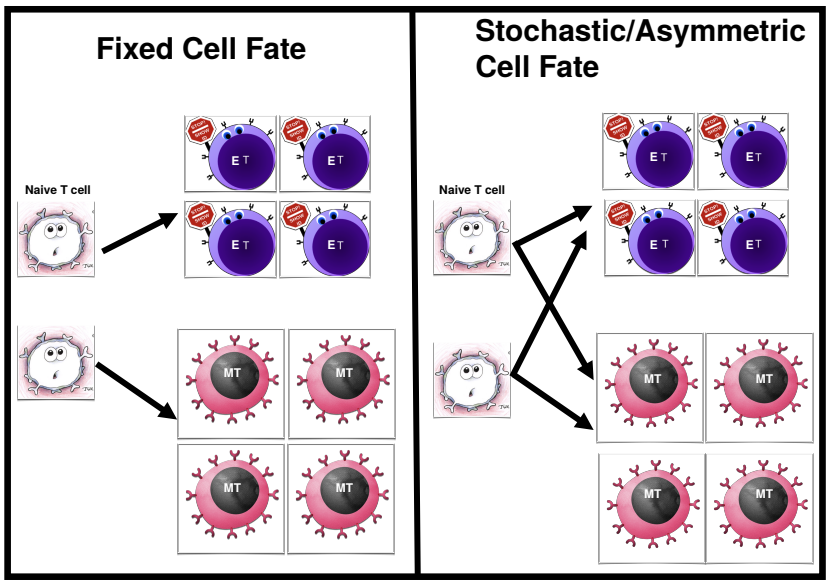

Figure 6.1: Contrasting scenarios of fixed and stochastic/asymmetric cell fate.

Two important observations that are common to [233, 234] are of heterogeneity in family sizes and heterogeneity in marker distribution. The contribution of the individual T cells at the peak of the immune response is highly variable: the largest family is about twice as large as the average or median. The type of marker expressed also serves to differentiate the families of cells, with smaller families having a large proportion of cells that are positive for the CD62L marker, and negative for KLRG-1, whereas in the larger families, more cells are positive for KLRG-1 and negative for CD62L. There is essentially no correlation between family size and CD27. Fig. 6.2 that has been adapted from [233] presents this data from the experimental study.

6.2.1 The Branching Process Model

The fate of a cell, namely the time to division or death, is determined by internal factors. A given cell will therefore either divide into two daughter cells at particular time, or will die. If the former event occurs, the factors in the parent cell are stochastically distributed to the daughters, which then independently decide their fate. On the other hand, if the individual cells are allowed to expand, in time they build a population which can then

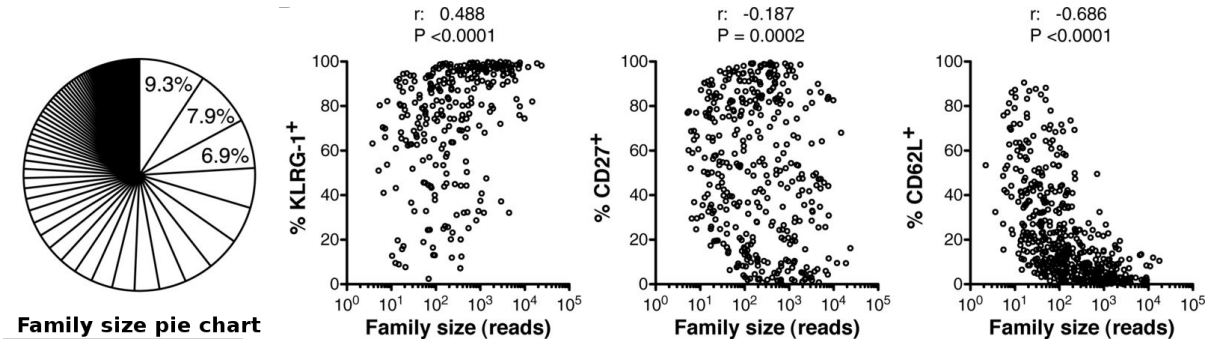

Figure 6.2: Adapted from Gerlach et.al. [233], and presented here as providing the motivation for the model being proposed.

be monitored for various behaviours.

I have expanded upon a model defined by Markham et.al [236] where correlations in B cell populations were ascribed to the result of a branching process model. This model has two factors: the propensity to divide (R factor) and the time to death (T factor), and these are stochastically inherited at each cell division. Two thresholds are defined for R: if the cell contains R above an upper threshold, it will definitely divide, while if it is below a lower threshold, it will die. In the intermediate range the fate is proportion to the amount of R. Three other factors, X, Y and Z were introduced with different rates of accumulation, and these are expected to mimic the behaviour of markers at the population level. The mathematical description of the model is as follows.

$$R_{p+1} = R_d + ((R_p + R_{m-d})/2) \tag{6.1}$$

$$T_{p+1} = T_p k_d \tag{6.2}$$

$$X_{p+1} = (X_p + (T_p k_x))x_d \tag{6.3}$$

$$Y_{p+1} = (Y_p + (T_p k_y))y_d \tag{6.4}$$

$$Z_{p+1} = Z_p z_d \tag{6.5}$$

$$k_x < \text{mean } T < k_y \tag{6.6}$$

Subscripts define the generation, here p and d indicate parent and daughter respectively. There is an additional element of stochasticity, with R_d being the variation added in daughter, while R_{m-d} is variation added in the mother cell. The parameters k_d, x_d, y_d are the rates of accumulation of respective factors. The constants are defined such that on average factor X is diluted, factor Y is concentrated and factor Z is randomly distributed in the population.

6.2.2 Simulation and Results

The above model was simulated using Matlab [237] starting with all factors being either uniformly or normally distributed. The effect of initial variation in the population is minor and the results are presented for a moderate variation at the start. The population of the T cells grows as shown in Fig. 6.3. The data is curated so that only the cells surviving at the peak population are considered for the rest of the analysis. In Fig. 6.4a the size of each family is represented as a sector in the pie diagram. Fig. 6.5 is the histogram of families sizes: there are a few large families and many small families. With an arbitrary cut-off, all the cells are defined positive or negative for factors depending on if they have crossed the cut-off. Figure 6.4 plots percentage of positive cells for factor X, Y and Z by size of the family. The percentage positives for factor X is negatively correlated with the family size and is positively correlated for factor Y. The correlations is absent for factor Z.

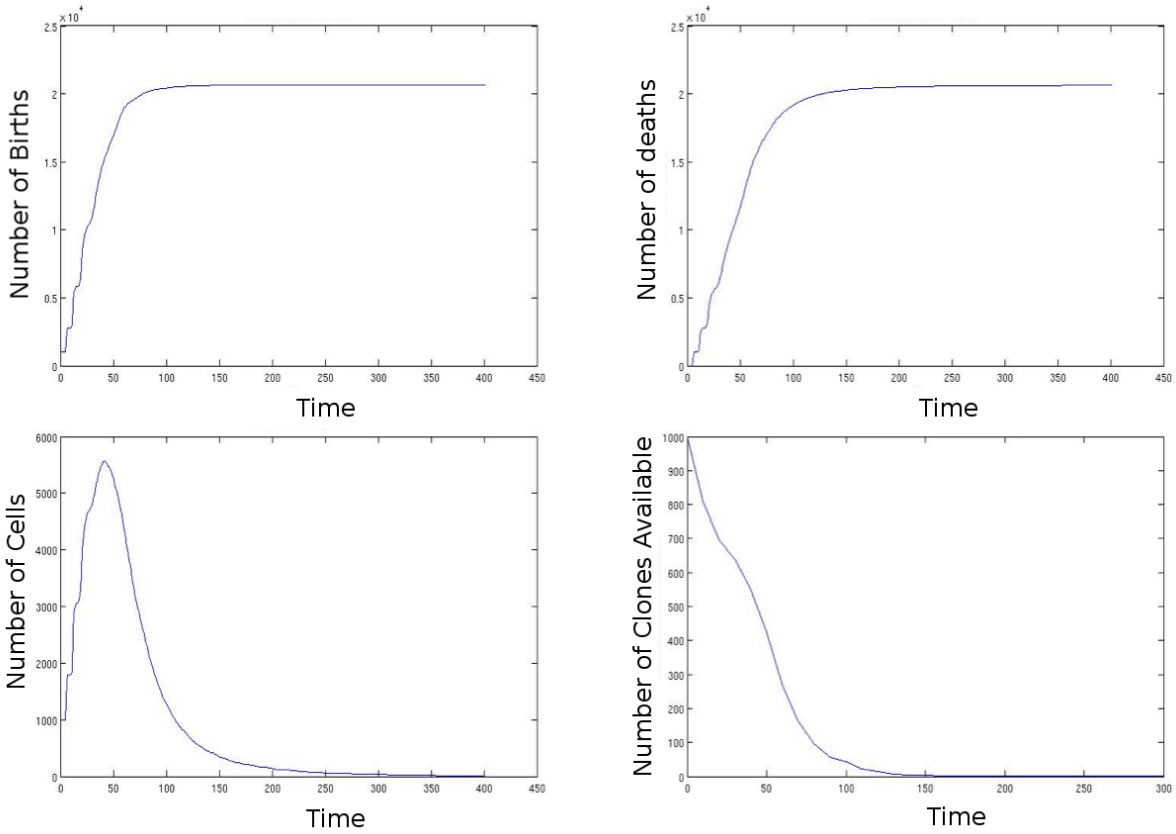

Figure 6.3: Time variation of different population attributes.

The experiments [233, 234] have also shown that if the organism is reinfected, the frequency distributions in the population is reproducible. There is a strong correla-

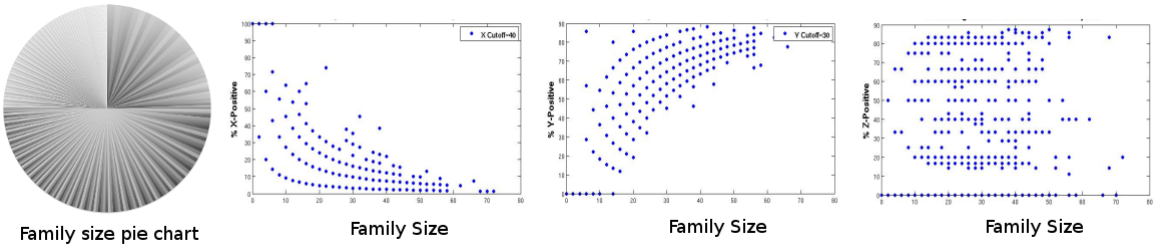

Figure 6.4: Pie diagram for family sizes at the peak response and percentage positives for factors X, Y, Z. Compare the results of this model with Fig. 6.2.

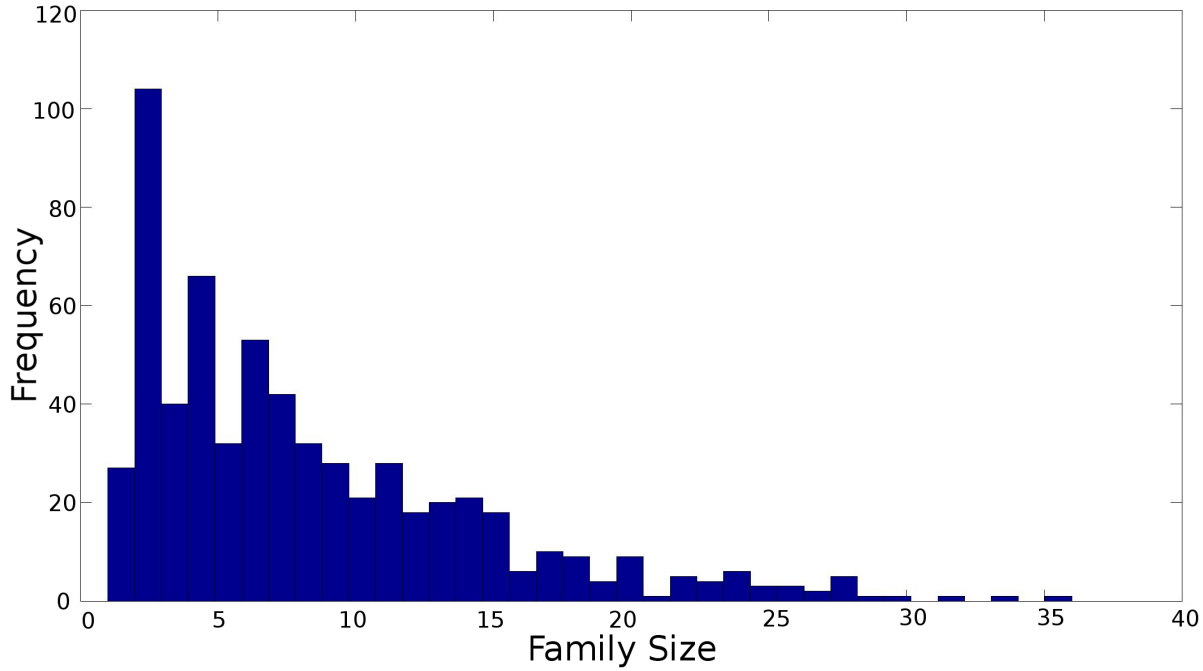

Figure 6.5: Histogram of family size frequencies.

tion between the family sizes for first and second infection [233]. To understand such behaviour, I discuss a generalised branching process model which can maintain the frequency distribution. To start with, the branching process model has a single internal factor X which controls cell division time τ and is produced at rate α. I assume the simplest dependence for τ, namely that it is always equal to X. In the following equations, the subscript is a time or generation index, while the superscript index labels the daughter cells.

$$\tau = f(X) \equiv X \tag{6.7}$$

$$X_1 = X_0 + \alpha(\tau) = X_0 + \alpha X_0. \tag{6.8}$$

If the factor is equally divided between the daughter cells, then each will have

$$X_1^1 = X_2^1 = 0.5(X_0 + \alpha X_0). \tag{6.9}$$

If α=1, then $X_1^1 = X_0$. Since $f(X) = X$, the cells with lower X divide faster than cells with larger X, leading to inverse relation between frequency of cells with X. Results of the simulation are shown in Fig. 6.6 and suggest that a simple single factor will not be able to maintain the frequency distribution of the population. A second factor is required.

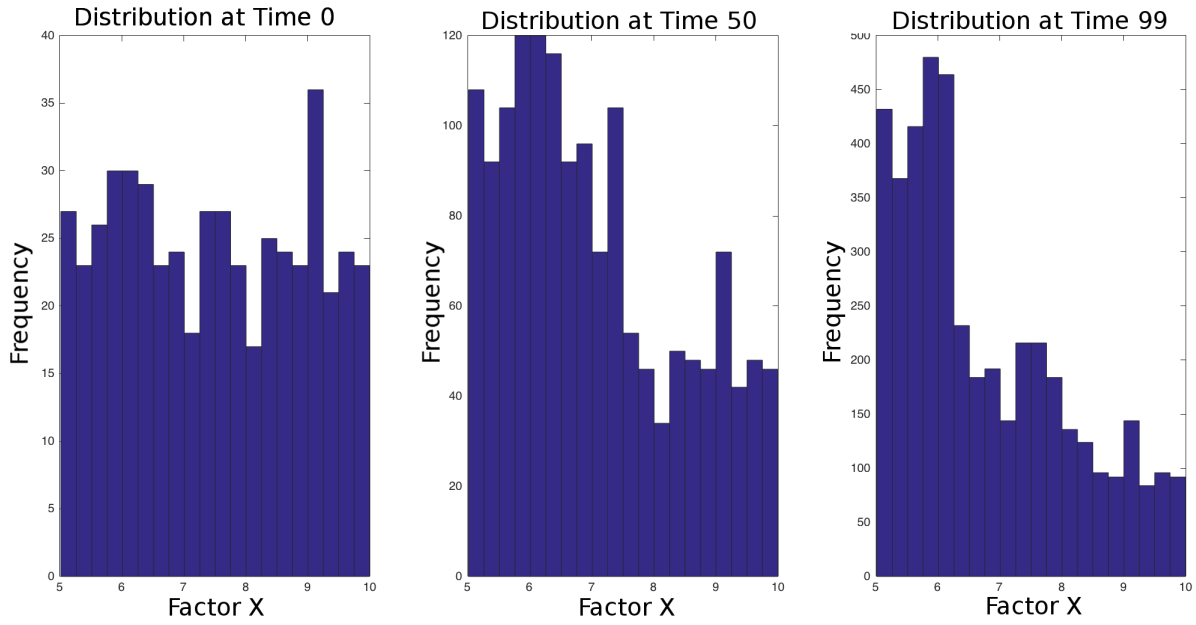

Figure 6.6: Frequency distribution of factor X in the evolving population at three time points in the single factor model.

Now assuming the existence of a second factor Y that has an anti-apoptotic effect and is produced at rate β, similar to X, namely

$$Y_1 = Y_0 + \beta(\tau) = Y_0 + \beta X0 \tag{6.10}$$

If this is equally divided between daughter cells then each will have.

$$Y_1^1 = Y_2^1 = 0.5(Y_0 + \beta X_0) \tag{6.11}$$

If $\beta = 1$ then $Y_0 = X_0$, namely Y_0 approaches tX_0 after a few cell divisions. The results of the simulation shown in Fig. 6.7 suggests that the two factor will be able to maintain the frequency distribution of the population. However, further analysis and simulations are needed to verify the robustness of the results.

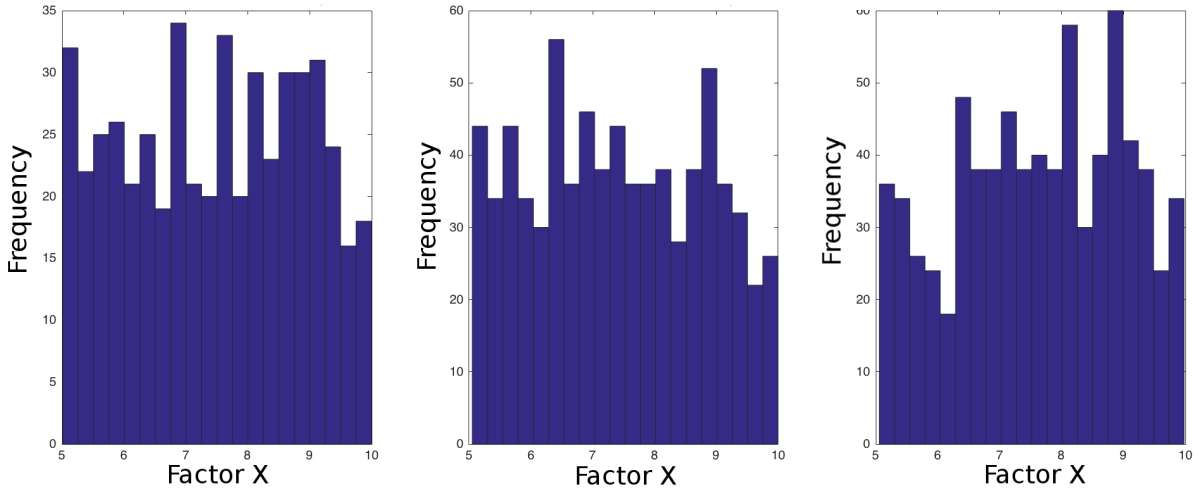

Figure 6.7: Frequency distribution of factor X in the evolving population at three time points in the two factor model; cf. Fig. 6.6.

6.2.3 Observations and Conclusions

This model shows a consistent behaviour for the size of population as a function of time. Family sizes show heterogeneity, with few large families and many smaller families. The factors X and Y show opposite correlations with family size. The population expanded from this model qualitatively matches the experimental results, although the extreme heterogeneity of family sizes evident in the experimental results cannot be replicated in the model. It is likely that multiple factors affect the time of division in a multiplicative manner and a model that includes such behaviour may provide some clues.

I have studied how population structure evolves due to branching process, and demonstrated that independently distributed inheritance factors can have specific behaviour at the population level. The R and T factors which decide fate of cell to divide generated a structure in the population and this structure along with inheritance properties of the independent factors X, Y, and Z gives the observed behaviours. In the generalised branching process model, I have shown that two simple sub-cellular factors were required

to maintain the frequency distribution of a population. This work was done in Prof. Rob de Boer's group (University of Utrecht) during a four month research exchange visit as part of Marie Sklodowska-Curie Actions of European Commission.

6.3 Protein-protein interaction networks: Random forest method

Proteins most often work cooperatively in order to perform specific biological functions. Proteins interact with each other through high specific physicochemical connections, electrostatic interactions and van der Waals forces. Collectively, these are termed protein-protein interactions (PPI) [238]. PPIs are fundamental to processes such as cellular localisation, signal transduction, cellular adhesion, inflammation and pathogen invasion. Deciphering PPIs can improve our understanding about the concerned biological process.

The so–called interactome is the network representation of PPIs. This provides a systems level framework to understand cellular, physiological and pathological processes [239]. Host-pathogen interactions are one area wherein protein-protein interactions are of great utility. PPIs play an important role in invasions, establishment and survival of pathogens. Several examples are known [240, 241, 242]. It is also generally believed that the construction and analysis of host-pathogen PPI networks may help in detecting drug targets and designing effective therapies [243, 244].

Important experimental methods used to decipher the host-pathogen PPIs are two hybrid screening [245, 246, 247, 248] and mass spectrometry coupled affinity purification [249]. Experimentally verified host-pathogen PPI networks (interactomes) are available for several viruses [249, 248, 246, 247] and bacteria [250, 251, 252, 253]. The information is available though several public databases such as the Human Immunodeficiency Virus (HIV)-1 Human-Interaction Database [254], HPIDB [255], VirusMentha [256], and Denvlnt [257]. Although reliable, these experimental methods are resource and time consuming. In addition the large number of false negatives limits their utility [258, 259]. Computational methods can therefore play a complementary role in host-pathogen PPIs identification [258, 260]; these include homology based [261, 262], domain-domain interactions [263], and machine learning-based methods [258, 264].

We have constructed a PPI network using [264] for the system of Human –*Helicobacter pylori*, a corkscrew shaped gram-negative bacterium that is present in the stomach and duodenal lining of half the world's population. This is the most common bacterial infection of man [265] and is associated with large number of gastric diseases such as gastritis [266], duodenal ulcer[267], gastric cancer [265] and intestinal metaplasia [268]. The biology of *H. pylori* infection, colonisations, immune evasion, survival and pathogenicity are still incompletely understood [269]. As discussed above the deciphering of host-pathogen

protein-protein interactions (PPIs) and and building of a network can provide insight into such processes.

6.3.1 Methods

The Multi-scale Local Descriptor-Random Forest (MLD-RF) approach for predicting protein-protein interactions from their sequences [264] consists of three steps,

1. Creation of training data by selecting and curating protein sequences for positive and negative datasets and gathering the unknown (test) protein sequence pairs,
2. Feature extraction and MLD representation of the protein sequence pairs, and
3. training a Random Forest ensemble classifier and applying it to unknown protein pairs.

We follow the procedure elaborated by You et al. [264] to generate positive and negative sets for the training data. The positive set consists of experimentally known interacting protein-protein pairs, curated to remove protein sequences with fewer than 50 amino acid residues, and protein pairs with more than 40 percent sequence identity. It is difficult to establish the lack of interaction between protein pairs, and therefore the negative set is generated assuming that most protein pairs show no interaction: all the possible protein pairs are considered as a "pre-negative" set, from which known interacting and/or co-localising protein pairs are removed. Depending on the size of the positive set an equal number of non-interacting protein pairs are selected from the curated pre-negative set to form a negative set. The method [264] is allegedly robust to which species constitute the training set. Due to lack of experimentally determined human-*H. pylori* PPI dataset and the robustness of the model, we assumed that the *H. pylori* training set can be used as a proxy. All the protein sequences (larger than 50 amino acid residues) from Human and *H. pylori* are arranged into all possible protein pairs. If n and m are the number of proteins from Human and *H. pylori* respectively, then total host-pathogen protein pairs will be equal to $n \times m$: this is the test set of sequences.

MLD is used to generate features from the above positive, negative and test set sequences. On each scale the protein sequence is divided into q sub-sequences of equal length and represented by q-bit binary string. For example, if the total sequence is divided into q=3 sub-sequences, these are denoted 000, 001, 010, 011, 100, 101, 111 where 0 or 1 denotes exclusion or inclusion of the particular sub-sequence. The number of such new continuous sequences will be $2^q - 2$, with q being varied between 4 and 7.

Amino acids can be grouped into p=7 classes depending on volumes of the side chains and the dipoles. For each continuous sequence the amino acids are replaced by the group number and are characterised by three types of descriptors: composition (C), transition (T) and distribution (D). The percentage of each type of amino acid gives its composition (C) and is a vector of length equal to number of groups of amino acids p. The transition frequency (percentage) of one group into another, given by T, is a vector of length $p \times (p-1)$. In a sequence the first, 25%, 50%, 75%, and 100% location of each group amino acids is given by D, a vector of length $4 \times p$. A new vector combining C, T and D forms the descriptor for each continuous sequence. The vector combing the descriptors of all the continuous sequences of a protein sequence pair forms its particular Multi-Scale Local Feature Representation.

As described in Chapter 1, Random forest clarifier (RFC) is an ensemble method with tree depth and number of trees as prominent parameters to be tuned. We used the MLD of the training set to train the RFC and classified the test set pair into PPI or not PPI. The positive and negative dataset consisted of 2986 data points each.

6.3.2 Implementation and results

To check the accuracy of the Random forest classifier, we conducted a 5-fold cross validation on the training set of *H. pylori* PPIs. The accuracy of RFC for our training data is 89.4%. As mentioned above the *H. pylori* PPI trained RFC is assumed to successfully classify *Helicobacter pylori* and *Human* PPIs. We have considered a subset of Human and *H. pylori* proteins for this analysis. Since *H. pylori* infection is mostly confined to the gastrointestinal tract [270], we consider only the proteins specific to human GI tract including stomach. There are 625 such proteins that can be identified in the online database of the Human Protein Atlas project [271]. The secretome of *H. pylori*, which included the set of all its secretory proteins, are the proteins most likely to interact with human proteins and so we considered only these for the analysis. All possible pairs of *H. pylori* secretome proteins (a total of 165) and human GI tract specific proteins (625) are envisaged, giving a total of 103125 possible pairs. The features of these protein pairs are extracted and the *H. pylori* PPI trained RFC is used to classify each pair into either PPI or not.

This procedure resulted in the network shown in Fig. 6.8, where each node is a protein and an edge indicates an interaction. The method is expected to be liberal in finding PPIs, overestimating the number of interactions, but will miss fewer true interactions. On the contrary, experimental methods like the yeast two-hybrid are very conservative, and are believed to underestimate the number of true interactions. Considering this, the network that we have generated is dense with many PPIs; see Fig. 6.8.

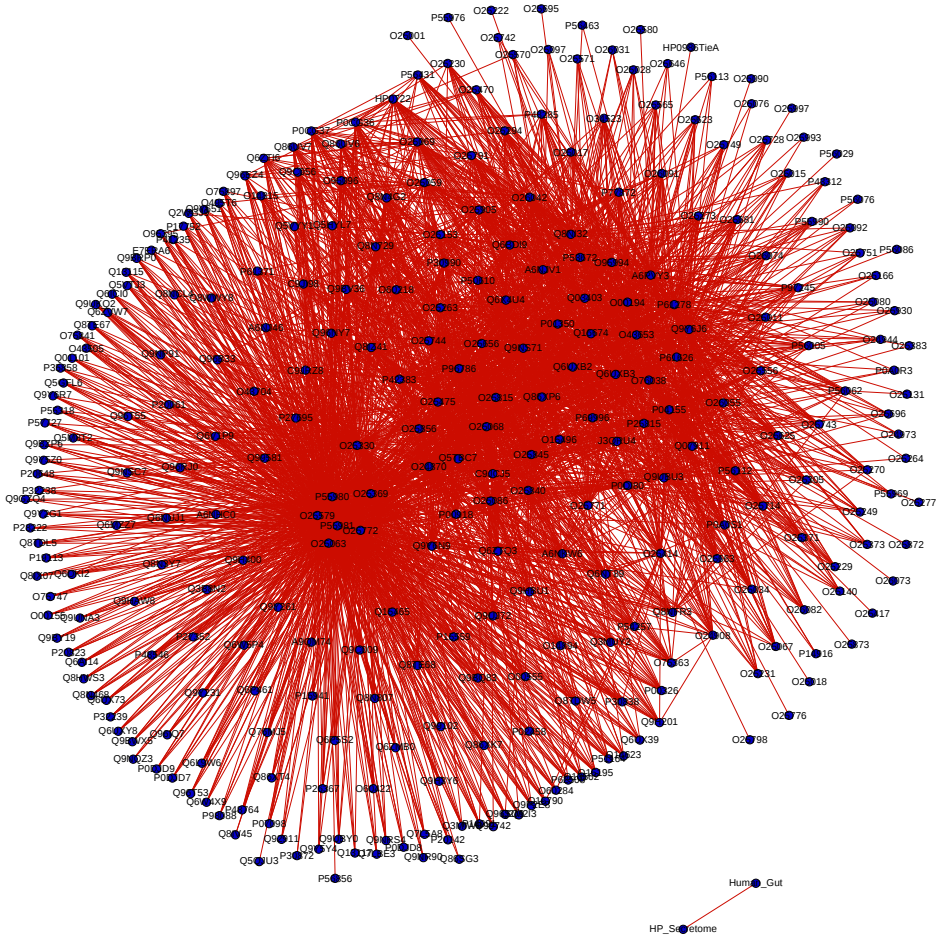

Figure 6.8: The *H. pylori* secretome, namely the human gut specific protein - protein interaction network.

The model unfortunately fails to predict the experimentally known interaction between the proteins HopQ and CEACAM, a known interaction. This is possibly since we are trying to predict eukaryotic PPI learning from prokaryotic interactions. To check the same we have tried predicting the known human-human PPIs and find that the pre-

dictive power of the model is low, approximately 60%. To improve the predictability of the model we need to train the model with human PPIs or with another eukaryote like *S. cerevisiae*. As the secretome of *H. pylori* will be modified due to protein digestion enzymes of Human stomach. There is a strong possibility that this can lead to miss-classification of PPIs.

6.3.3 Observations and Conclusion

The multi-scale local-descriptor Random Forest approach makes it possible to finding putative host-pathogen PPIs. The method is robust in finding the PPIs of the species whose data is used to train the RFC, but cross-species application leads to low accuracy, suggesting that a more inclusive training data is required. In the future I plan to extend this work so as to include PPIs from diverse sources such as eukaryotes, prokaryotes, cross species PPIs etc. in the training data and then determine the accuracy when applied to different known PPIs. If the model is found to be robust, the *H. pylori* and human PPI network will be generated for further analysis. I also wish to identify the characteristic protein digestion sites of human stomach proteases, and further use this data to computationally generate modified protein sequences of *H. pylori* secretome. These sequences might help in better identification of PPIs. This work is being done in collaboration with Dr. Ramani and Sumeet Tiwari.

Bibliography

[1] A. B. Novikoff, Science, **101**, 209–215, (1945).

[2] I. Lobo, Nature Education Knowledge, **1**, 141 (2008).

[3] M. H. Wake, Integr. Comp. Biol., **43**, 23–241 (2003).

[4] R. Levins and R. Lewontin, *The Dialectical Biologist* (Harvard university Press, USA, 1985).

[5] S. W. Robinson, M. Fernandes and H. Husi, Comput. Struct. Biotechnol. J., **11**, 35–46 11.

[6] A. Prokop and B. Csukas (eds.), *Systems Biology: Integrative Biology and Simulation Tools* (Springer, Netherlands, 2013).

[7] G. W. Brodland, Semin. Cell Dev. Biol., **47-48**, 62–73 (2015).

[8] H. P. Fischer, Alcohol Res. Health., **31**, 49–59 (2008).

[9] R. D. Huntoon and A. Weiss J. Res. Natl. Bur. Stand., **38**, 39–410 (1947).

[10] E. T. Liu, Cell, **121**, 505–506 (2005).

[11] J. D. Murray, Mathematical Biology: 1. An Introduction (Springer, New York, 2004).

[12] A. J. M. Garrett, *Ockham's Razor*. In: W. T. Grandy and L. H. Schick(eds) *Maximum Entropy and Bayesian Methods. Fundamental Theories of Physics*,(Springer, Dordrecht, 1991).

[13] A. J. Lotka, J. Phys. Chem., **14**, 271–274 (1910).

[14] V. Volterra, Atti Congr. intern. dei Mat. a Bologna, **1**, 215–232 (1928).

[15] A. M. Turing, Phil. Transact. Royal Soc. B, **237** 37–72 (1952).

[16] A. L. Hodgkin and A. F. Huxley, J Physiol., **117**, 500–544 (1952).

[17] R. May, *Stability and Complexity in Model Ecosystems*, (Princeton university Press, Princeton and Oxford, 1973).

[18] S. Wolfram, Rev. Mod. Phys., **55**, 601–644 (1983).

[19] M. A. Nowak. *Evolutionary Dynamics: Exploring the equations of life*, (Harvard University Press, Cambridge, 2006).

[20] M. Mitchell, *An Introduction to Genetic Algorithms*, (MIT Press, Cambridge, 1996).

[21] M. A. Du Plessis, Oecologia, **90**, 205–211 (1992).

[22] T. R. Malthus, *An Essay on the Principle of Population*, (London, 1798).

[23] P. F. Verhulst, Corresp. Math. Phys. **10**, 113–121 (1838).

[24] D. Ruelle, IHES Publ. Math. **50**, 27 (1979).

[25] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617–656 (1985).

[26] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, 1993).

[27] A. M. Lyapunov, *On the stability of ellipsoidal forms of equilibrium of rotating fluids (in Russian), Master's dissertation*, University of St. Petersburg, 1884.

[28] M. Cencini, F. Cecconi and A. Vulpiani, *Chaos: From Simple Models to Complex Systems*, (World Scientific, Singapore, 2010).

[29] K. Ramasubramanian and M. S. Sriram, Physica D: Nonlinear Phenomena, **139**, 72–86 (2000).

[30] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering* (Perseus Books, New York, 1994).

[31] A. Prasad, S. S. Negi, and R. Ramaswamy, Int. J. Bifurcation Chaos, **11**, 291-309 (2001).

[32] J. D. Crawford, Rev. Mod. Phys. **63**, 991 (1991).

[33] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications* (Springer-Verlag, New York, 1976).

[34] A. N. Pisarchik and U. Feudel, Phys. Rep. **540**, 167 (2014).

[35] J.C. Alexander, J.A. Yorke, Z You, and I. Kan, Int. J. Bifurcation Chaos Appl. Sci. Eng. **02**, (1992) 795–813

[36] E. Ott, J. C. Sommerer, J. C. Alexander, I. Kan, and J. A. Yorke, Physica D **76**, 384-410 (1994).

[37] M.L. Heltberg, R.A. Kellogg, S. Krishna, S. Tay, and M.H. Jensen, Cell Syst. **3**, 532–539 (2016).

[38] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2003)

[39] S. R. Ujjwal, N. Punetha, A. Prasad, and R. Ramaswamy, Phys. Rev. E **95**, 032203 (2017). T Wontchui, J. Y. Effa, H. P. E Fouda, S. R. Ujjwal, and R. Ramaswamy, Phys. Rev. E **96**, 062203 (2017).

[40] J. M. Smith, Mathematical Ideas in Biology (Cambridge University Press, New York, 1968).

[41] R. C. Lewontin, J. Theor. Biol., **1**, 382–403 (1961).

[42] J. W. Weibull, *Evolutionary Game Theory* (MIT Press, 1997)

[43] J. M. Smith, Physica D: Nonlinear Phenomena, **22**, 43–49 (1986).

[44] M. A. Nowak, Science, **303**, 793–799 (2004).

[45] M. J. Osborne. *An Introduction to Game Theory*, (OUP, USA, 2003)

[46] R. Andino and E. Domingo, Virology, **479–480**, 46–51 (2015).

[47] L. P. Villarreal, and G. Witzany, World. J. Biol. Chem., **4**, 79 (2013).

[48] G. Szabó, and G. Fáth, Phys. Rep., **446**, 97–216 (2007).

[49] M. A. Nowak, C. E. Tarnita, and E. O. Wilson, The evolution of eusociality, Nature, **466**, 1057–1062 (2010).

[50] S. M. Sanchez and T. W. Lucas, Proceedings of the Winter Simulation Conference, 116–126 (2002).

[51] P. Davidsson, J. Holmgren, H. Kyhlbäck, D. Mengistu, and M. Persson, Applications of Agent Based Simulation in Multi-Agent-Based Simulation VII, (Springer-Verlag, Berlin, 2007).

[52] C. M. Macal and M. J. North, J. Simulat., **4**, 151–162 (2010).

[53] C. M. Macal, and M. J. North, Proceedings of the 2009 Winter Simulation Conference, 86–98 (2009).

[54] L. H. Encinas, S. H. White, A. M. del Rey, and G. R. Sánchez, Appl. Math. Model., **31**, 1213–1227 (2007).

[55] W. V. Winkle, H. I. Jager, S. F. Railsback, B. D. Holcomb, and T. K. Studley, Ecol. Model., **110**, 175–207 (1998).

[56] L. Stein, Nat. Rev. Genet., **2**, 493–503 (2001).

[57] J. Pevsner, Bioinformatics and Functional Genomics (Third) (Wiley Blackwell, New Jersey, 2015).

[58] J. J. Pasternak, An Introduction to Human Molecular Genetics (Second) (John Wiley and Sons, New Jersey, 2005)

[59] H. P. J. Buermans, J. T. den Dunnen, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, **1842**, 1932–1941 (2014).

[60] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice M-A Rajandream and B. Barrell, Bioinformatics, **16**, 944–945 (2000).

[61] M. A. Lesk, Introduction to Genomics (Oxford University Press, Oxford, 2007).

[62] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, Nucleic Acids Res., **28**, 33–36 (2000).

[63] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B, Kiryutin, E./ V. Koonin, and D. A. Natale, BMC Bioinformatics, **4**, 1–14 (2003).

[64] NCBI-Resource-Coordinators, Nucleic Acids Res., **42**, D7–D17 (2014).

[65] V. M. Markowitz, I. M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N. N. Ivanova and N. C. Kyrpides, Nucleic Acids Res., **42**, D560–D567 (2014).

[66] M. Kimmel and Axelrod, *Branching Processes in Biology* (Springer 2002).

[67] K. S. Mode, Crump and J. Charles, J. Appl. Prob., 6, 205–210 (1969).

[68] H. W. Watson and Francis Galton, J. Anthropol. Inst. Great Brit., **4**, 138–144 (1875).

[69] R. Bellman and T. E. Harris, Proc. Natl. Acad. Sci. Unit. States Am., **34**, 601–604 (1948).

[70] O. Hyrien, R. Chen and M. S. Zand, Biology Direct, **5**, 41 (2010).

[71] O. Hyrien, M. M. Pröschel, M. Noble and A. Yakovlev, Biometrics, **61**, 199–207 (2005).

[72] A. Yates, C. Chan, J. Strid, S. Moon, R. Callard, A. J. T. George and J. Stark, BMC Bioinformatics, **8**, 1–20 (2007).

[73] C. Gerlach, J. C. Rohr, L. Perié, N. V. Rooij, J. W. J. V. Heijst, A. Velds and T. N. M Schumacher, Science, **340**, 635–639 (2013).

[74] Carmen Molina-ParÃs and Grant Lythe, Mathematical Models and Immune Cell Biology (Springer, London, 2011)

[75] E. D. Hawkins, J. F. Markham, L. P. McGuinness and P. D. Hodgkin, Proc. Natl. Acad. Sci. Unit. States Am., **106**. 13457–13462 (2009).

[76] E. J. Wherry, V. Teichgräber, T. C. Becker, D. Masopust, S. M. Kaech, R. Antia and R. Ahmed, Nat. Immunol., **4**, 225–234 (2003).

[77] R. S. Day, Math. Biosci., **78**, 73–90 (1986).

[78] D. Wick and S. G. Self, Math. Biosci., **165**, 115–134 (2000).

[79] C. J. Mode1, C. K. Sleeman and T. Raj, Front. Genet., **4**, 1–11 (2013).

[80] S. Shalev-shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press, New York, 2014).

[81] T. M. Mitchell, Machine learning (McGraw-Hill, New York, 1997).

[82] J. B. MacQueen, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, **1**,281–297 (1967).

[83] J. A. Hartigan and M. A. Wong, J. Roy. Stat. Soc. C Appl. Stat., **28**, 100–108 (1979).

[84] L. Kaufman and P. J. Rousseeuw, Finding groups in data: An introduction to cluster analysis (Wiley, New York, 1990)

[85] S. C. Johnson, Psychometrika, **32**, 241–254 (1967).

[86] T. Kondo, Proceedings of the 37th SICE annual conference, IEEE, 1143–1148 (1998).

[87] S. Fine, Y. Singer and N. Tishby, Mach. Learn., **32**, 41–62 (1998).

[88] V. Vapnik, The support vector method of function estimation, in J. A. K. Suykens and J. Vandewalle (Eds) Nonlinear Modeling: Advanced Black-Box Techniques (Kluwer Academic Publishers, Boston, 1998).

[89] D. D. Lewis, Naive (Bayes) at forty: The independence assumption in information retrieval. in C. Né dellec, C. Rouveirol (eds) Machine Learning: Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence) (Springer, Berlin, Heidelberg, 1998).

[90] C. Kingsford and S. L. Salzberg, Nat. Biotechnol., **26**, 1011–1013 (2008).

[91] J. R. Quinlan, Mach. Learn., **1**, 81–106 (1986).

[92] L. Breiman L, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees (Wadsworth International Group, Belmont, CA, USA, 1984).

[93] R.E. Schapire. The boosting approach to machine learning: an overview. In: D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, and B. Yu, Nonlinear Estimation and Classification (Springer, New York, 2003).

[94] L. Breiman. Mach. Learn., **45**, 5–32 (2001).

[95] A. Liaw and M. Wiener, R news, **2**, 18–22 (2002).

[96] R. J. Trudeau, Introduction to Graph Theory (Dover Publications, Mineola, New York, 1993).

[97] R. Diestel, Graph Theory (Springer-Verlag, New York 1997).

[98] D. J. Watts and S. H. Strogatz, Nature, **393**, 440–442 (1998).

[99] T. McGlynn, Nature Education Knowledge, **3**, 69 (2010).

[100] E. O. Wilson and B. Holldobler, Proc. Natl. Acad. Sci. Unit. States Am., **102**, 13367–13371 (2005).

[101] B. J. Crespi and D. Yanega, Behav. Ecol. **6**, 109–115 (1995).

[102] R. Gadagkar, Current Science, **64**, 215–216 (1993).

[103] L. Keller, Queen Number and Sociality in Insects (Oxford Scientific, New York, 1993).

[104] L. Keller and M. Genoud, Lett. to Nat., **389**, 958–960 (1997).

[105] M. Kleiber, The Fire of Life: An Introduction to Animal Energetics (Wiley: New York, 1961).

[106] G. G. Carter and G. S. Wilkinson, Proc. R. Soc. B, **280**, 20122573 (2013).

[107] R. A. Bradshaw and E. A. Dennis, Handbook of Cell Signaling, (Cell Biology) (Academic Press, Burlington, MA, 2004).

[108] F. Mercurio and A. M. Manning, Oncogene, **18**, 6163–6171 (1999).

[109] L. Partridge and N. H. Barton, Nature, **362**, 305–311 (1993).

[110] P. B. Medawar, Modern Q, **1**, 30–56 (1946).

[111] M. A. Rose, Evolutionary Biology of Aging (Oxford University Press, Oxford, 1991)

[112] T. B. L. Kirkwood, Nature, **270**, 301–304 (1997).

[113] V. Gohil and A. Joshi, Resonance, **3**, 67–72 (1998).

[114] J. Heinze and A. Schrempf, Gerontology, **54**, 160–167 (2008).

[115] L. D. Mueller and M. R. Rose, Proc. Natl. Acad. Sci. Unit. States Am. **93**, 15249–15253(1996).

[116] P. B. Medawar, An unsolved problem of biology (Lewis, London, 1952).

[117] W. D. Hamilton, J. Theor. Biol., **12**, 12–45 (1996).

[118] B. Charlesworth, Genetics, **156**, 927–931 (2000).

[119] C. L. Rauser, J. J. Tierney, S. M. Gunion, G. M. Covarrubias, L. D. Mueller and M. R. Rose, J. Evol. Biol, **19**, 289–301 (2006).

[120] M. R. Rose, Evolution **38**, 1004–1010 (1984).

[121] Z. Bas, R. Bijlsma and R. F. Hoekstra, Evolution, **49**, 649–659 (1995).

[122] H. A. Kimberly and M. R. Rose, Annu. Rev. Entomol., **50**, 421–445 (2005).

[123] D. N. Reznick, M. J. Bryant, D. Roff, C. K. Ghalambor and D. E. Ghalambor, Nature., **431**, 1095–1099 (2004).

[124] H. Y. Chen and M. A. Maklakov, Curr. Biol., **22**, 2140–2143 (2012).

[125] M. N. Shokhirev and A. A. Johnson, PLoS ONE, **9**: e86602 (2014).

[126] Y. L. Cunff, A. Baudisch and K. Pakdaman, PLoS Comput. Biol., **9**, e1002825 (2013).

[127] B. H Kramer and R. Schaible, PLos ONE, **8** (2013).

[128] P. Nonacs, Evolution, **42**, 566–580 (1988).

[129] L. Keller, Tree, **10**, 355–360 (1995).

[130] E. O. Wilson and B. Holldobler, the Ants (Springer, Berlin, 1990).

[131] A. F. G Bourke, N. R Franks, Social Evolution in Ants (Princeton University Press, Princeton, 1995).

[132] B. H. Kramer and R. Schaible, Biol. J. Linn. Soc., **109**, 710–724 (2013).

[133] X. Liao, S. Rong and D. C. Queller, PLOS Biology, **13**, 1–14 (2015).

[134] E. ". Charnov, R. Warne, and M. Moses, Am. Nat., **170**, E129–E142 (2007).

[135] P. H. Leslie, Biometrika, **33**, 183–212(1945).

[136] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

[137] J. C. Thiele, J. Stat. Software., **58**, 1–41 (2014).

[138] U. Wilensky, NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston. URL: http://ccl.northwestern.edu/netlogo/.

[139] P. A. Abrams, Evolution, **47**, 877–887 (1993).

[140] M. Reichard, Semin. Cell Dev. Biol., **70**, 99–107 (2017).

[141] A. A. Maklakov, L. Rowe and U. Friberg, BioEssays, **37**, 802–807 (2015).

[142] A. A. Maklakov and S. Immler, Curr. Biol., **26**, R577-R586 (2015).

[143] M. Hammers, S. A. Kingma, K. Bebbington, J. van de Crommenacker, L. G. Spurgin, D. S. Richardson, T. Burke, H. L. Dugdale and Komdeur J, Exp. Gerontol., **71**, 69–79 (2015).

[144] R. Blažek, M. c. Polačik, P. Kačer, A. Cellerino, R. Řežucha, C. Methling, O. Tomášek, K. Syslová, E. T. Terzibasi, T. Albrecht, M. Vrt´ilek and M. Reichard, Evolution, **71**, 386–402 (2015).

[145] M. I. Lind, H-Y. Chen, S. Meurling, A. C. G. Guevara, H. Carlsson, M. K. Zwoinska, J. Andersson, T. Larva and A. A. Maklakov, Funct. Ecol., **6**, 1252–1261 (2017).

[146] E. M. L. Duxbury, W. G. Rostant and T. Chapman, Proc. Biol. Sci., **284**, 20170391 (2017).

[147] H. Ancell and A. Pires-daSilva, Semin. Cell Dev. Biol., **70**, 122–129 (2017).

[148] R. D. Lee, Proc. Natl. Acad. Sci. Unit. States Am., **100**, 9637-9642 (2003).

[149] L Tesfatsion, Artif. Life, **8**, 55–82 (2002).

[150] G. An, Q. Mi, J Dutta-Moscato and Y Vodovotz, Wiley Interdiscip. Rev. Syst. Biol. Med., **1** 159–71 (2009).

[151] E. M. Joshua, Complexity, **4**, 41–60 (1999).

[152] V. Grimm and S. F. Railsback, Individual-Based Modeling and Ecology (Princeton University Press, Princeton and Oxford, 2005).

[153] A. Lomnicki, Population Ecology from the Individual Perspective *Individual-Based Models and Approaches Ecology: Populations, Communities and Ecosystem* ed. DeAngelis D L and Gross L J (Chapman and Hall, New York, 1992).

[154] R. Axelrod, The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration (Princeton University Press, Princeton, 1997)

[155] M. W. Macy and R. Willer, Annu. Rev. Sociol., **28**, 143–66 (2002).

[156] G. G. Carter and G. S. Wilkinson, Proc. R. Soc. B, **282**, 2015–2524 (2015).

[157] S. A. West, S. A. Griffin and A. Gardner, Curr. Biol., **17**, R661–R672 (2007).

[158] A. Grafen, Natural selection, kin selection and group selection in Behavioural Ecology: An Evolutionary Approach, 2nd Edition ed J. R. Krebs and N. B. Davies (Blackwell Scientific Publications, Oxford, 1984).

[159] L. Lehmann and L. Keller, J. Evol. Biol. **19**, 1365–76 (2006).

[160] W. Hamilton, J. Theor. Biol., **7**, 1-16 (1964).

[161] R. C. Connor, Phil. Trans. R. Soc. B, **365**, 2687–97 (2010).

[162] G. S. Wilkinson, Nature, **308**, 181-184 (1984).

[163] B. K. McNab, J. Mammal., **54**, 131–144 (1973).

[164] G. S. Wilkinson, Ethol. Sociobiol., **9**, 85–100, (1988).

[165] G. S. Wilkinson, Anim. Behav., **34**, 1880–1889 (1986).

[166] M. Witkowski, Adapt. Behav., **15**, 307–328 (2007).

[167] A. M. Greenhall, G. Joermann and U. Schmidt, Desmodu rotundus (Mammalian Species), The American Society of Mammalogists, **202**, 1-6, (1983).

[168] Clutton-Brock T 2009 Cooperation between non-kin in animal societies *Nature* **462** 51–57

[169] C. M. Gray, P. König, A. K. Engel and W. Singer, Nature **338**, 334–337 (1989).

[170] A. Hastings and T. Powell, Ecology **72**, 896–903 (1991).

[171] R. Lev Bar-Or, R. Maya, L. A. Segel, U. Alon, A. J. Levine and M. Oren, Proc. Natl. Acad. Sci. Unit. States Am. **97**, 11250–11255 (2000).

[172] L. Stone and D. He, J. Theor. Biol. **248**, 382–390 (2007).

[173] F. Jülicher and J. Prost, Phys. Rev. Lett. **78**, 4510–4513 (1997).

[174] A. Hoffmann, A. Levchenko and M. L. Scott, Science **298**, 1241–1245 (2002).

[175] A. T. Winfree, *The Geometry of Biological Time* (Springer-Verlag, New York, 2001)

[176] J. C. Dunlap, Cell **96**, 271–290 (1999).

[177] A. Nandi, C. Vaz, A. Bhattacharya and R. Ramaswamy, BMC Syst. Biol. **3**, 1–16 (2009).

[178] M. S. Hayden and S. Ghosh, Genes Dev. **26**, 203–234 (2012).

[179] Q. Zhang, M. J. Lenardo and D. Baltimore, Cell **168**, 37–57 (2017).

[180] A. A. Beg and D. Baltimore, Science **274**, 782–784 (1996).

[181] L. A. J. O'Neill and C. Kaltschmidt, Trends Neurosci. **20**, 252–258 (1997).

[182] P. A. Baeuerle and T. Henkel, Annu. Rev. Immunol. **12**, 141-179 (1994).

[183] T. Lawrence, Cold Spring Harb. Perspect. Biol. **1**, 1–10 (2009).

[184] U. Feudel, Int. J. Bifurcation Chaos, **18**, 1607-26 (2008).

[185] S. R. Ujjwal, N. Punetha, R Ramaswamy, M Agrawal and A Prasad, Chaos **26**, 063111 (2016).

[186] S. H. Strogatz, Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering, (Perseus books, New York, 1994)

[187] L. M. Pecora, F. Sorrentino, A. M. Hagerstrom, T. E. Murphy and R. Roy, Nat. Commun. **5**, 4079 (2014).

[188] S. Krishna, M. H. Jensen and K. Sneppen, Proc. Natl. Acad. Sci. Unit. States Am., **103**, 10840–10845 (2006).

[189] M. H. Jensen and S. Krishna, FEBS Lett. **586**, 1664–1668 (2012).

[190] M. L. Heltberg, S. Krishna and M. H. Jensen, J. Stat. Phys. **167**, 792–805 (2017).

[191] G. Olmos and J. Lladó, Mediators of Inflamm. **2014**, 1–12 (2014).

[192] W.-M Chu, Canc. Lett., **328**, 222–225 (2012).

[193] S.I. Grivennikov, A.V. Tumanov, D.J. Liepinsh, A.A. Kruglov, B.I. Marakusha, A.N. Shakhov, T. Murakami, L.N. Drutskaya, I. Förster, B.E. Clausen, L. Tessarollo, B. Ryffel, D.V. Kuprash and S.A. Nedospasov, Immunity, **22**, 93–04 (2005).

[194] E.E Varfolomeeva and A. Ashkenazi, Cell, **116**, 491–497 (2004).

[195] M. Karin, Oncogene, **18**, 686–6874 (1999).

[196] M. Hinz and C. Scheidereit, EMBO Rep. **15**, 46–61 (2014).

[197] B. Mengel, A. Hunziker, L. Pedersen, A. Trusina, M. H. Jensen and S. Krishna, Curr. Opin. Genet. Dev. **20**, 656–664 (2010).

[198] M. H. Jensen, L. P. Kadanoff and S. Krishna, preprint (2017).

[199] R. A. Kellogg and S. Tay, Cell **160**, 381–92 (2015).

[200] H. E. Nusse and J. A. Yorke, Science **271**, 1376–1380 (1996).

[201] J. Huisman and F. J. Weissing, Am. Nat. **157**, 488–494 (2001).

[202] A. Daza, A. Wagemakers, B. Georgeot, D. Guery-Odelin and M. A. F. Sanjuan, Sci. Rep. **6**, 31416 (2016).

[203] P. Ashwin and O. Burylko, Chaos **25**, 013106 (2015).

[204] T. Y. Tsai, Y. S. Choi, W. Ma, J. R. Pomerening, C. Tang and J. E. Ferrell Jr., Science **321**, 126–129 (2008).

[205] S. Zambrano, I. DeToma, A. Piffer, M. E. Bianchi and A. Agresti, eLife **5**, e09100 (2016).

[206] D. Angeli, J. E. Ferrell Jr. and E. D. Sontag, Proc. Natl. Acad. Sci. Unit. States Am., **101**, 1822-1827 (2004).

[207] E. Ullner, A. Zaikin, E. I. Volkov and J. Garcia-Ojalvo, Phys. Rev. Lett. **99**, 148103 (2007).

[208] R. Li and B. Bowerman, Cold Spring Harb. Perspect. Biol., **2**, a003475 (2010).

[209] M. M. Markiewski and J. D. Lambris, Am. J. Path. **171**, 715–727 (2007).

[210] S. Halstenberg, A. Panitch, S. Rizzi, H. Hall and J. A. Hubbell, Biomacromolecules **3**, 710–723 (2002).

[211] J. Allen, J. Theor. Biol., **165**, 609–631, (1993).

[212] D. L. Theobald, Nature, **465**, 219–222 (1995).

[213] M. Lynch, J. S. Conery, Science, **302**, 1401–1404 (2003).

[214] M. Lynch, The Origins of Genome Architecture (Sinauer, Sunderland, MA, 2007).

[215] J. B. S. Haldane, On being the right size (Simon and Schuster, New York, 1956).

[216] W. Alonso, Pap. Reg. Sci. Assoc., **26**, 67–83, (1971).

[217] N. W. Pirie, Annu. Rev. Microbiol., **27**, 119–132 (1973).

[218] E. V. Koonin and Y. I Wolf, Nucleic Acids Res., **36**, 6688–6719 (2008).

[219] N. Lane and W. Martin, Nature, **467**, 929–934 (2010).

[220] R. L. Tatusov, M. Y. Galperin, D. A. Natale and E. V. Koonin, Nucleic Acids Res., **28**, 33–36 (2000).

[221] E. V. Nimwegen, Trends Genet., **19**, 479–484 (2003)

[222] L. J. Croft, J. L. Martin, M. J. Gagen and J. S. Mattick, Genome Biology, **5**, (2003).

[223] M. J. Gagen and J. S Mattick, Theory Biosci., **123**, 381–411 (2005).

[224] G. Beslon, D. P. Parsons, Y. Sanchez-Dehesa, J. M. Pena and C. Knibbe, Biosystems, **102**, 32–40 (2010).

[225] J. B. Arnold, BioEssays, **6**, 279–282 (1987).

[226] N. Lane, Biol. Direct, **6**, 35, (2011).

[227] T. R. Gregory, Nat. Rev. Genet., **6**, 699–708 (2005).

[228] K. T. Konstantinidis and J. M. Tiedje, Proc. Natl. Acad. Sci. Unit. States Am., **101**, 3160–3165 (2004).

[229] E. V. Koonin, Proc. Natl. Acad. Sci. Unit. States Am., **112**, 15777–15778 (2015).

[230] R. J. Banavar, M. E. Moses, J. H. Brown, J. Damuth, A. Rinaldo, R. M. Sibly and A. Maritan, Proc. Natl. Acad. Sci. Unit. States Am., **107**, 15816-15820 (2010).

[231] T. J. Kindt, R. A. Goldsby, B. A. Osborne and J. Kuby, Kuby immunology, (W.H. Freeman, New York, 2013).

[232] E. J. Wherry, V. Teichgräber, T. C. Becker, D. Masopust, S. M. Kaech, R. Antia, U. H. von Andrian and R. Ahmed, Nat. Immunol., **340**, 225–234 (2003).

[233] C. Gerlach, J. C. Rohr, L. Perié, N. van Rooij, J. W. J. van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, R. J. de Boer and T. N. M. Schumacher, Science, **340**, 635–639 (2013).

[234] V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer and D. H. Busch, Science, **340**, 630–635 (2013).

[235] o. T. Chang, V. R. Palanivel, I. Kinjyo1, F. Schambach, A. M. Intlekofer, A. Banerjee, S. A. Longworth, K. E. Vinup, P. Mrass, J. Oliaro, N. Killeen, J. S. Orange, S. M. Russell, W. Weninger and S. L. Reiner, Science, **315**, 1687–1691 (2007).

[236] J. F. Markham, C. J. Wellard, E. D. Hawkins, K. R. Duffy and P. D. Hodgkin, J. Royal Soc. Interface., **7**, 1049–1059 (2010).

[237] MathWorks, MATLAB - Version 7.10.0, (2010). `http://www.mathworks.com`

[238] J. D. L. Rivas and C. Fontanillo, PLoS Comput. Biol., **6**, e1000807 (2010).

[239] M. E. Cusick, N. Klitgord, M. Vidal and D. E. Hill, Hum. Mol. Genet. **14**, R171–81 (2005).

[240] A. Choy, J. Dancourt, B. Mugo, T. J. O'Connor, R. R. Isberg, T. J. Melia and C. R. Roy, Science, **338**, 1072–1076 (2012).

[241] N. Dong, Y. Zhu, Q. Lu, L. Hu, Y. Zheng and F. Shao, Cell, **150**, 1029–1041 (2012).

[242] N. Noinaj, N. C. Easley, M. Oke, N. Mizuno, J. Gumbart, E. Boura, A. N. Steere, O. Zak, P. Aisen, E. Tajkhorshid, R. W. Evans, A. R. Gorringe, A. B. Mason, A. C. Steven and S. K. Buchanan, Nature, **483**, 53–58 (2012).

[243] Y. Feng, Q. Wang, T. Wang, Biomed. Res. Int., **2017**, 1289259 (2017).

[244] M. Skwarczynska and C. Ottmann, Future Med. Chem., **7** 2195–219 (2015).

[245] K. H. Young, Biol. Reprod., **58**, 302–311 (1998).

[246] S. D. Shapira, I. Gat-Viks, B. O.V. Shum, A. Dricot, M. M. de Grace, L. Wu, P. B. Gupta, T. Hao, S. J. Silver, D. E. Root, D. E. Hill, A. Regev and N. Hacohen, **139**, 1255–1267 (2009).

[247] Z. H. Davis, E. Verschueren, G. M. Jang, K.Kleffman, J. R. Johnson, J. Park, J. V. Dollen, M. C. Maher, T. Johnson, W. Newton, S. Jäger, M. Shales, J. Horner, R. D. Hernandez, N. J. Krogan and B. A. Glaunsinger, Mol. Cell, **57**, 349–360, (2015).

[248] P. Uetz,Y. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum and J. Haas, Science, **311**, 239–42 (2006).

[249] S. Jäger et.al., Nature, **481**, 365–370 (2012).

[250] M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali and B. W. Sobral, PLoS ONE, **5**, e12089 (2010).

[251] H. Yang, Y. Ke, J. Wang, Y. Tan, S. K. Myeni, D. Li, Q. Shi, Y. Yan, H. Chen, Z. Guo, Y. Yuan, X. Yang, R. Yang and Z. Du, Infect. Immun., **79**, 4413–4424 (2011).

[252] V. Memiŝević, N. Zavaljevski, R. Pieper, S. V. Rajagopala, K. Kwon, K. Townsend, C. Yu, X. Yu, D. DeShazer, J. Reifman and A. Wallqvist, Mol. Cell Proteomics, **12**, 3036–3051 (2013).

[253] S. Blasche, S. Arens, A. Ceol, G. Siszler, M. A. Schmidt, R. Häuser, F. Schwarz, S. Wuchty, P. Aloy, P. Uetz, T. Stradal and M. Koegl, Sci. Rep., **4**, 7531 (2014).

[254] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt and R. G. Ptak, Nucleic Acids Res., **37**, D417–D422 (2009).

[255] R. Kumar and B. Nanduri, BMC Bioinformatics, **11**, S16 (2010).

[256] A. Calderone, L. Licata and G. Cesareni, Nucleic Acids Res., **43**, D588–D592 (2015).

[257] L. Dey and A. Mukhopadhyay, PLoS Negl. Trop. Dis., **11**, e0005879 (2017).

[258] T. Cui, W. Li, L. Liu, Q. Huang and Z. He, PLoS ONE, **11**, e0147612 (2015).

[259] M. D. Dyer, T. M. Murali and B. W. Sobral, Bioinformatics. **23**, i159–66 (2007).

[260] E. Nourani, F. Khunjush and Saliha Durmuş, Front. Microbiol., **6**, 94,(2015).

[261] N. Tyagi, O. Krishnadev and N. Srinivasan, Mol. Biosyst., **5**,1630–1635 (2009).

[262] B. de Chassey, L. Meyniel–Schicklin, A. Aublin–Gex, V. Navratil, T. Chantier, P. André and V. Lotteau, EMBO reports, **14**, 938–994 (2013).

[263] X. Liu, Y. Huang, J. Liang, S. Zhang, Y. Li, J. Wang, Y. Shen, Z. Xu and Y. Zhao, BMC Bioinformatics, **15**, 393 (2014).

[264] Z. You, K. C. C. Chan and P. Hu, PLoS ONE, **10**, e0125811, (2015).

[265] M. Amieva and R. M. Peek, Gastroenterology, **150**, 64–78 (2016).

[266] D. C. Metz, H. C. Weber, M. Orbuch, D. B. Strader, I. A. Lubensky and R. T. Jensen, Dig. Dis. Sci., **40**, 153–159 (1995).

[267] E. J. Kuipers, J. C. Thijs and H. P. Festen, Aliment. Pharmacol. Therapeut., **9**, 59–69 (1995).

[268] K. S. Liu, I. O. Wong and W. K. Leung, World J. Gastroenterol., **22**, 1311–1320 (2016).

[269] J. G. Kusters, A. H. M. van Vliet and E. J. Kuipers, Clin. Microbiol. Rev., **19**, 449–490 (2006).

[270] S. Suerbaum and P. Michetti, N. Engl. J. Med., **347**, 1175-1186 (2002).

[271] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling and F. Ponten, Nat. Biotechnol., **28**, 1248–1250 (2010).